Towards more robust and accurate computations of capillary effects in the simulation of multiphase flows in porous media

Sabrina Bassetto

École Doctorale MADIS–631
UNIVERSITÉ DE LILLE

Vers une prise en compte plus robuste et précise des effets capillaires lors de simulations d'écoulements multiphasiques en milieux poreux

Towards more robust and accurate computations of capillary effects in the simulation of multiphase flows in porous media

Thèse préparée à
Inria Lille-Nord Europe, IFP Energies nouvelles et Laboratoire Paul Painlevé

et soutenue publiquement par

Sabrina Bassetto

le 16/12/2021, à Rueil-Malmaison, pour obtenir le grade de Docteur en
Mathématiques Appliquées

devant le jury composé de

Claire Chainais-Hillairet	Professeure des universités, Université de Lille	Présidente
Anna Scotti	Professeure associée, Politecnico di Milano	Rapporteur
Mazen Saad	Professeur des universités, École Centrale de Nantes	Rapporteur
Danielle Hilhorst	Directeur de recherche, Université Paris-Saclay	Examinatrice
Konstantin Brenner	Maître de conférences, Université Côte d'Azur	Examinateur
Clément Cancès	Directeur de recherche, INRIA Lille-Nord Europe	Directeur
Guillaume Enchéry	Ingénieure de recherche, IFP Energies nouvelles	Co-encadrant
Quang-Huy Tran	Ingénieure de recherche, IFP Energies nouvelles	Co-encadrant

Acknowledgements

Ce manuscrit représente l'aboutissement d'un parcours très enrichissant du point de vue scientifique et humain. C'est pourquoi je tiens à remercier en premier lieu les personnes qui ont contribué à cette expérience. Premièrement, je souhaite remercier mon directeur de thèse Clément Cancès pour m'avoir accompagnée durant ces trois années. Merci pour les fois où toi et Clémence m'avez chaleureusement accueillie chez vous lors de mes séjours à Lille. Merci d'avoir été toujours disponible pour m'aider, clarifier mes doutes et m'expliquer les choses que je ne savais pas. Et merci pour la confiance que tu as placée en moi. Mes remerciements vont ensuite à mon co-directeur de thèse Quang-Huy Tran pour l'attention accordée à mes travaux et sa sympathie. Une personne incroyable, omnisciente, à laquelle je porte un profond respect. Merci également à Guillaume Enchéry pour m'avoir prise en stage et m'avoir encadrée durant cette thèse. Je tiens à t'exprimer mes plus sincères remerciements pour ton implication durant ce parcours.

J'adresse ensuite mes remerciements à Claire Chanais-Hillairet pour avoir accepté de présider mon jury de thèse ainsi que pour les moments partagés à Lille ! Je suis également reconnaissante envers Anna Scotti et Mazen Saad pour avoir accepté le rôle de rapporteurs et pour leur appréciation de mes travaux. In particolare grazie ad Anna per essere venuta di persona ad assistere alla discussione, mi ha fatto molto piacere. Je tiens aussi à exprimer mes remerciements à Danielle Hilhorst et Konstantin Brenner pour avoir participé au jury en tant qu'examinateurs.

Au cours de ces trois années à IFPEN, j'ai eu le plaisir de rencontrer des personnes spéciales, collègues et doctorants, que je tiens à remercier. Tout d'abord je souhaite remercier Zakia Benjelloun-Touimi pour m'avoir accueillie au sein du département de Mathématiques Appliquées. Ensuite, mes pensés vont aux thésards. Merci à Bastien et Karine qui m'ont accueillie à mon arrivée en m'intègrant dans le groupe des thésards et qui sont devenus pour moi des chers amis. Merci pour votre soutien, pour m'avoir écoutée quand j'en avais besoin, pour les bons moments que nous avons passés ensemble et pour votre amitié qui, j'en suis sure, durera. Merci à mes compagnons d'aventure : bg Guissel, Alexis le sage et Thoi la solaire. Nous avons affronté chaque étape de ce voyage ensemble, en partageant des moments difficiles et en célébrant nos réalisations ensemble. Merci à Ruben pour nos conversations et les soirées entre Italiens qui m'ont permis de me sentir chez moi. Merci à Joëlle et Karim, merci pour les bons moments passés ensemble, pour nos conversations et nos rires. Enfin, merci à Julien, Zakariae et Jin : c'était un plaisir de vous rencontrer ! Je vous souhaite le meilleur pour la réalisation de vos projets et beaucoup de bonheur !

Je tiens également à remercier tous mes collègues avec qui j'ai pu partager de bons moments en dehors du travail. Je pense à Bruno L., un ami sur lequel on peut compter. Merci pour nos belles et interminables conversations, pour avoir été mon collègue de dégustation, pour ta convivialité et ta gentillesse à mon égard. Merci à Frédéric N., que j'ai rencontré lors des cours d'œnologie et avec qui j'ai eu le plaisir de partager cette passion commune pour le vin. Merci à Delphine et Christian dont la convivialité et la bonne humeur ont animé les moments que nous avons passés ensemble, me donnant sourires et rires. Merci à Sylvie P. pour le soutien que tu m'as apporté au fil des

années, tu as su comprendre mes moments de difficulté et m'aider à les surmonter avec discrétion et délicatesse. Et que dire de mes collègues de bureau, les meilleurs ! Benjamin qui, malgré son aura de sérieux, est une personne compréhensive et gentille. C'était un plaisir de te rencontrer et je te remercie pour nos conversations. Et puis il y a Francesco, mon compatriote. C'est une personne au cœur tendre, gentille, patiente et serviable. Merci de m'avoir aidée à résoudre mes doutes mathématiques et d'avoir supporté mes divagations, et merci pour les moments que nous avons partagés en dehors du bureau. Tu es un ami !

Je tiens également à remercier tous les autres collègues du département avec lesquels j'ai eu des échanges intéressants et que je n'oublierai certainement pas. Merci donc à Rodolphe, Abir, Jean-Yves, Jean-Louis, Nina, les deux Thomas, Thibault, Michel G., Frédéric D., Nicolas P. et Chakib B.

Je tiens également à remercier toutes les personnes que j'ai rencontrées lors de l'ASIP, des sorties de la CE et autour de la commission week-end, avec laquelle j'ai eu le plaisir de collaborer. Je pense à Frédéric M. et Estelle, Michel P., Nicole e Philippe, Véronique W., Sylvie et Lionel, Catherine et Dominique, Denis et Florence, Iryna. Merci à vous tous !

Grazie ad i miei amici più cari che, nonostante la lontananza, ci sono sempre stati. Grazie per i bei momenti passati insieme e per avermi sostenuta ed ascoltata quando ne avevo bisogno. Grazie quindi alla mia stellina Lucia, all'elegante Elia e all'irreperibile Salvo. Grazie ai miei amici Francesco e Rossella del gruppo di pianoforte che, se all'inizio è nato dal fatto che avevamo tutti la stessa insegnante, ora è un bellissimo gruppo di amici con cui condividere tanti momenti speciali. E in questo gruppo troviamo anche la nostra insegnante Silvia a cui rivolgo un ringraziamento particolare perchè per me è una persona speciale che porto sempre nel cuore.

Un pensiero affettuoso e un sincero ringraziamento alla mia sorellina Chiara ed ai miei genitori, Bertilla e Renato, che mi hanno sostenuta in questo percorso credendo in me. Nonostante in questi anni ci siano stati diversi kilometri a separarci, penso che questa distanza ci abbia reso più consapevoli ed aumentato l'amore che ci lega...anche se poi litighiamo comunque!! Ma come si suol dire "l'amore non è bello se non è litigarello". Un pensiero super affettuoso alla mia nonnina Giuse, che in questi anni mi ha scritto tante belle letterine e che, ogni volta che torno a casa, mi ricorda quanto amore provi per la sua nipotina lontana con le sue lacrime di gioia. Ti amo anche io nonnina.

E parlando di amore, quest'ultimo ringraziamento (ma non per importanza!) va a te tesoro. Grazie Francesco per essermi stato accanto. Sei stato la mia roccia, mi hai sopportata quando ero davvero insopportabile e mi hai saputo spronarmi nei momenti di sconforto. Grazie per tutto il tuo amore!

Summary

Carbon dioxide capture, utilization and storage (CCUS) is a powerful technology to reduce the quantity of greenhouse gases emitted into the atmosphere. Generally, CO_2 is stored in geological underground structures such as depleted oil and gas reservoirs or saline aquifer. Once injected into formations, CO_2 is trapped underground by means of various trapping mechanisms. The formation heterogeneities and changes in wettability are involved in one of them. The discontinuities thus created are at the basis of the capillary barrier phenomenon, which plays a crucial role for flows in porous media and in fractured ones in particular.

For Darcy flows, capillary pressure is often modeled as a function of fluid saturation and rock type. Each lithology corresponds to a capillary pressure-saturation curve which displays strong variations embodied by asymptotes. The change of curve induced by the change of rock requires to define precisely the interface conditions between two different lithologies in order to model the flow or the trapping of the fluids accurately through this interface. In view of these characteristics and constraints, numerical difficulties may arise when simulating these flows, especially during Newton iterations. Some choices of primary variables may be more appropriate than others.

In this thesis, we aim at improving Newton robustness in order to overcome the above-mentioned difficulties and at proposing strategies to enforce transmission conditions at interfaces in heterogeneous domains. Our work follows an order of increasing difficulties. First, we start considering the easier model, the Richards equation, in a homogeneous medium. Then, we introduce heterogeneities in the domain. Finally, we turn to the complete model in a challenging configuration: the immiscible incompressible two-phase system in a heterogeneous domain.

To improve robustness, we propose a strategy based on variable switch and it is easily implemented thanks to a fictitious variable that enables us to describe both the saturation and the pressure and that we call parametrization technique. The numerical tests performed confirm the potentiality of this technique, which allows the Richards equation to be solved without caring about the choice of the primary unknown and without any convergence problems.

In a heterogeneous domain, a naive scheme without explicit inclusion of heterogeneities suffers from a lack of accuracy in the predicted results. This motivates the introduction of a specific treatment of the interfaces. Thus, we propose and compare several approaches to deal with the interface transmission condition, analyzing their pros and cons when confronted to different physical settings for the Richards equation as well as the two-phase Darcy flow model.

Résumé

La séquestration du dioxyde de carbone constitue une technologie puissante permettant de réduire la quantité de gaz à effet de serre émis dans l'atmosphère. En général, le CO_2 est stocké dans des structures géologiques souterraines telles que des réservoirs de pétrole et de gaz épuisés ou des aquifères salins. Une fois injecté dans les formations, le CO_2 est piégé dans le sous-sol au moyen de divers mécanismes de piégeage. Les hétérogénéités de la formation et les changements de mouillabilité sont impliqués dans l'un d'eux. Les discontinuités ainsi créées sont à la base du phénomène de barrière capillaire, qui joue un rôle crucial pour les écoulements en milieu poreux et en particulier dans les milieux fracturés.

Pour les écoulements de Darcy, la pression capillaire est souvent modélisée en fonction de la saturation du fluide et du type de roche. A chaque lithologie correspond une courbe pression capillaire-saturation qui présente de fortes variations matérialisées par des asymptotes. Le changement de courbe induit par le changement de roche nécessite de définir précisément les conditions d'interface entre deux lithologies différentes afin de modéliser précisément l'écoulement ou le piégeage des fluides à travers cette interface. Compte tenu de ces caractéristiques et contraintes, des difficultés numériques peuvent apparaître lors de la simulation de ces écoulements, notamment lors des itérations de Newton. Certains choix de variables primaires peuvent être plus appropriés que d'autres.

Dans cette thèse, nous cherchons à améliorer la robustesse de Newton afin de surmonter les difficultés mentionnées ci-dessus et à proposer des stratégies pour faire respecter les conditions de transmission aux interfaces dans des domaines hétérogènes. Notre travail suit un ordre de difficultés croissantes. Tout d'abord, nous commençons par considérer le modèle le plus simple, l'équation de Richards, dans un milieu homogène. Ensuite, nous introduisons des hétérogénéités dans le domaine. Enfin, nous nous tournons vers le modèle complet dans une configuration difficile : le système diphasique incompressible immiscible dans un domaine hétérogène.

Pour améliorer la robustesse, nous proposons une stratégie basée sur le changement de variable et elle est facilement mise en œuvre grâce à une variable fictive qui nous permet de décrire à la fois la saturation et la pression et que nous appelons technique de paramétrisation. Les tests numériques réalisés confirment la potentialité de cette technique qui permet de résoudre l'équation de Richards sans se soucier du choix de l'inconnue primaire et sans problème de convergence.

Dans un domaine hétérogène, un schéma naïf sans prise en compte explicite des hétérogénéités souffre d'un manque de précision dans les résultats. Ceci motive l'introduction d'un traitement spécifique des interfaces. Ainsi, nous proposons et comparons plusieurs approches pour traiter la condition de transmission aux interfaces, en analysant leurs avantages et inconvénients lorsqu'ils sont confrontés à différents paramètres physiques pour l'équation de Richards ainsi que pour le modèle d'écoulement diphasique de Darcy.

Contents

Chapter 1

Introduction

This work comes within the area of simulation of multiphase flows in porous media, with CO_2 storage as the primary application. We first describe this context in §1.1, laying emphasis on some physical processes of interest for the sequel, among which capillarity is the most notable.

We then present in §1.2 the two mathematical models that will be considered throughout this thesis. Far from being the most comprehensive ones from the point of view of physical effects, they are nevertheless sufficiently representative of the difficulties that arise after numerical discretization. The highlight of these difficulties and a review of the state of the art in §1.3 will allow us to state the objectives and to provide an outline of the manuscript in §1.3.3.3.

1.1 Environmental and technical issues of CO_2 storage

Global warming is a complex and non-negligible problem which is affecting the Earth and living creatures. Signed on December 12th 2015 by 196 Parties at COP 21 in Paris, the Paris Agreement on climate change became effective on November 4th 2016. Its goal is to limit global warming to 1.5°C, compared to pre-industrial levels. Different strategies have been adopted or are under evaluation as instruments to alleviate climate change, such as greener technologies (nuclear energy, wind energy etc...). Unfortunately, the renewable energy development is slow and the demand for fossil fuels in the world remains high, implying an increasing amount of greenhouse gases emitted into the atmosphere.

1.1.1 Modes of storage

The carbon dioxide capture, utilization and storage (CCUS) technology is a powerful instrument to reduce the quantity of greenhouse gases emitted into the atmosphere. Different CO_2 sequestration projects are in progress or in planning status in different parts of the world. The most notable among these are the Sleipner project in Norway [131], the Weyburn project in Canada and the In Salah project in Algeria. Moreover, diverse pilot-scale projects have also been carried out across the world. They consist in injecting small quantities of CO_2 into established formations for a small period of time. The first pilot-scale project located in USA is Frio [92]. These projects provide valuable information about the behavior of the carbon dioxide during the process. Also the field-scale injections of CO_2 have brought a greater understanding of the physics of the processes involved in storage and on the monitoring tools which could be used for large-scale injections. This need for information stems from the complexity and the risks of this process, such as carbon dioxide leakage and induced seismicity. Numerical simulations have been performed to ascertain

these risks [114, 134]. Moreover, the complexity of the problem should not be overlooked, which makes its modelling challenging. This is a multi-scale problem both in the temporal and spatial scales in which physical and chemical mechanisms are involved.

Before speaking about the modeling of CO_2 storage, let us provide more details about this process. Storage modes can be classified into natural and man-made modes. The former includes terrestrial sequestration (store the gas into soils and vegetation), while the latter includes storage in geological formations and is the most largely used in sequestration technology: CO_2 is stored in geological underground structures such as depleted oil and gas reservoirs or saline aquifers (Figure 1.1). Depleted oil and gas reservoirs are fields that have been classified as uneconomical for further production. These storage sites are already very well-known and there exist numerical computer models of these formations which have been validated, providing enhanced confidence in them. Moreover, these sites have been able to safely store oil and gas during a long time so they are the favorite candidates for storage. The existing infrastructures and wells can be used for CO_2 injection. On the other hand, the store capacity of these sites is lower than that of the saline aquifer formations because of the necessity to avoid excessively high pressures that can damage the cap rock. As already pointed out, saline aquifer formations are characterized by their highest storage capacity. For example, for Alberta deep saline basin, an estimate of the order of 103 Gt storage capacity has been made [87].

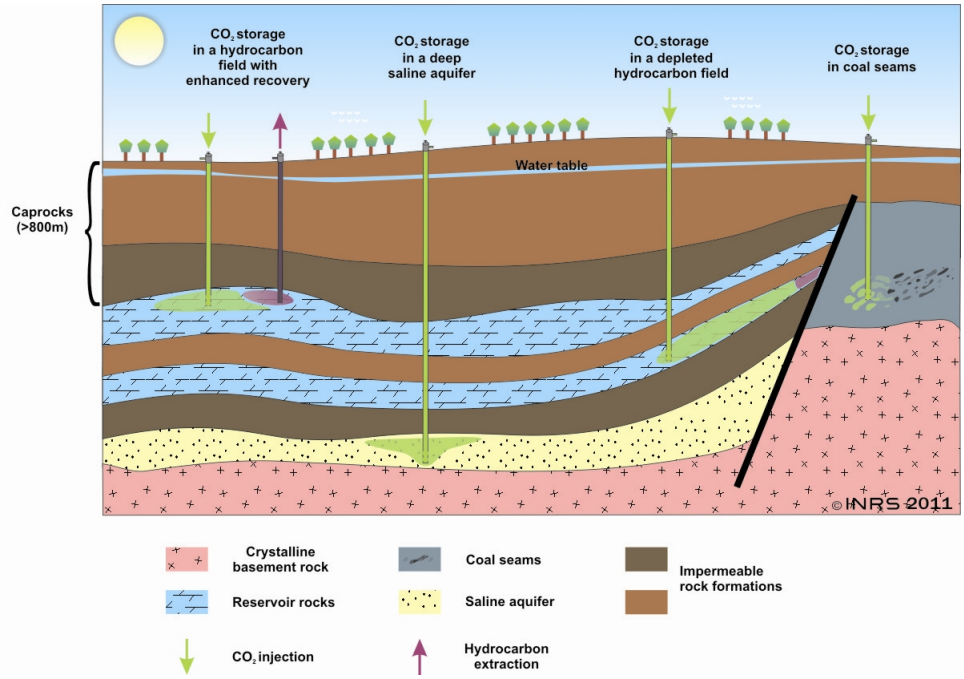


Figure 1.1: Different types of formations for geological storage of CO_2. `http://grrebs.ete.inrs.ca/wp-content/uploads/2014/02/stockage_eng.jpg`

Let us now discuss about storage process of CO_2 in geological formations with a focus on saline aquifers. A geological site to be selected as storage site, has to satisfy three main requisites: the capacity, the injectivity and the containement. The capacity constraint guarantees that the site has large pore volumes to store big quantities of CO_2, i.e., an high porosity. The porosity

$$\phi = \frac{\text{pore volume}}{\text{total volume}}$$

measures the fraction of the volume of voids over the total volume (Figure 1.2). If it also possesses a high permeability (denoted by λ), which is a measure of the ability of a porous material to allow fluids to pass through it, then the injectivity properties is satisfied, which ensures that lower wellhead pressures can be employed to preserve desired injection rates. To avoid that injected carbon dioxide leaks into groundwater or evades to the surface, because of the lower density of this gas with respect to resident brine, adequate cap rocks (rocks characterized by very low permeability) and sealing faults (if present) are needed. CO_2 is stored in a supercritical phase, i.e., it is compressed to higher pressures and temperatures about 89°F and 7.4 MPa. It reduces the buoyancy differential between CO_2 and present fluids.

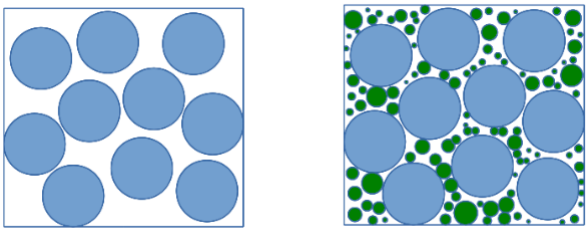


Figure 1.2: Example of media characterized by high and low porosities.

1.1.2 Physical processes involved

1.1.2.1 Trapping mechanisms

The injected supercritical CO_2 is then trapped underground by means of two main trapping mechanisms: physical trapping and geochemical trapping.

Physical trapping. In this category we can distinguish two types of trapping. One is the structural trapping, in which the formation heterogeneities and changes of wettability play a crucial role. Indeed an heterogeneous medium is characterized by different porosities, $\phi = \phi(x)$, and permeabilities, $\lambda = \lambda(x)$. This mechanism is similar to the one that has maintained oil and gas stored for millennia. When the injection ceases, the supercritical CO_2 tends to migrate upward via the porous and permeable rock because of the buoyancy effect, and laterally through preferential pathways, until it reaches a medium characterized by a low permeability cap rock, sealed discontinuities [89]. This prevents further migrations (Figure 1.3).

The second mechanism belonging to physical trapping is the residual/capillary trapping (Figure 1.4). While the carbon dioxide at supercritical state pervades the storage formations, reservoir fluids are displaced and they fill the remaining spots. CO_2 moves upward because of the density differences and laterally as effect of viscous forces. The CO_2 movement is stopped by the surface tension between CO_2 and brine [124]. Thereby CO_2 is trapped in the pores at residual gas saturation. This is called capillary effect.

The capillary pressure in a porous media depends on the wettability and interfacial tension changes [24]. It is defined as the difference between the pressures of the non-wetting and wetting phases. In a CO_2-water system, CO_2 is the non-wetting phase and water is the wetting one. Capillary pressure formula is given by the Young-Laplace equation

$$p_c = p_{CO_2} - p_w = \frac{2\sigma_{w,CO_2}\cos\theta}{R},$$

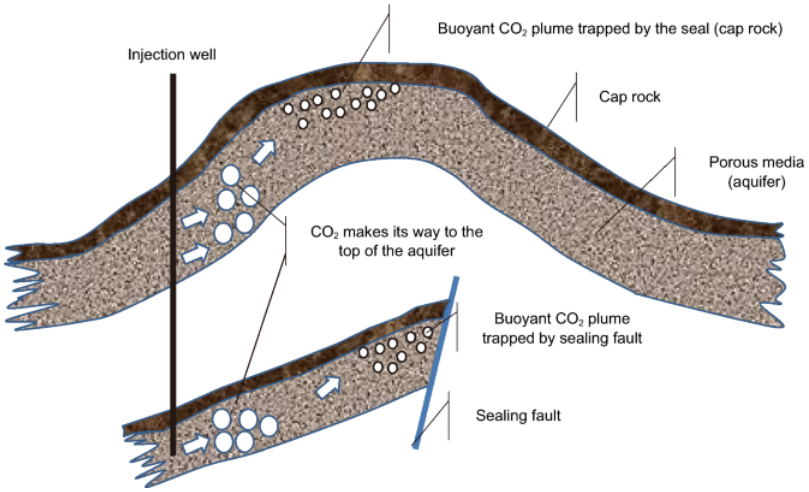


Figure 1.3: Physical trapping of the injected CO_2 due to the stratigraphic structure of the formation [86].

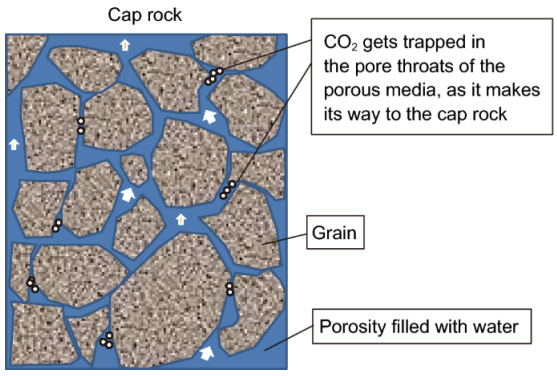


Figure 1.4: Residual trapping of CO_2 due to the pore structure of the formation. The CO_2 plume movement is indicated by the arrows [86].

where R is the pore radius, σ_{w,CO_2} is the interfacial tension between water and CO_2, and θ is the contact angles between the wetting medium and the rock surface (Figure 1.5).

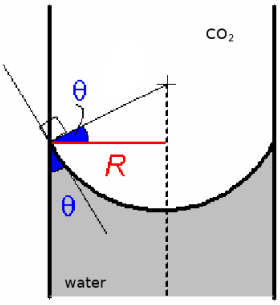


Figure 1.5: Young-Laplace's equation: definition of the capillary pressure in a tube.

The contact angle is the angle between the surface of the liquid and the outline of the contact surface and it is a measure of the wettability, of a solid by a liquid. Young's equation

$$\sigma_{sg} = \sigma_{sl} + \sigma_{lg} \cdot \cos\theta$$

describes the relationship between the contact angle θ, the surface tension of the liquid σ_{lg}, the interfacial tension σ_{sl} between liquid and solid and the surface free energy σ_s of the solid (Figure 1.6).

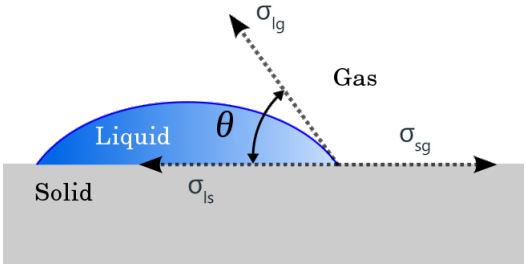


Figure 1.6: Schematic figure of contact angle.

The wettability of a fluid depends on its surface tension, the forces that drive a fluid's tendency to take up the minimal amount of space possible, and it is determined by the contact angle of the fluid. So this angle is crucial to determine which is the wetting and non wetting phase. If $\theta < 90°$, water is the wetting phase but if $\theta > 90°$ then CO_2 is the wetting phase (Figure 1.7).

Geochemical trapping. Geochemical trapping occurs when CO_2 reacts with the formation brine and the rock. CO_2 no longer appears as a separate phase, increasing storage capacity and promoting long-term storage. In this category we find the solubility trapping and the mineral trapping. The solubility trapping is the result of the CO_2 dissolution in the brine, leading to dense CO_2-saturated brine. An example of a possible reaction is

$$CO_{2(aq)} + H_2O \leftrightarrow H^+ + HCO_3^-.$$

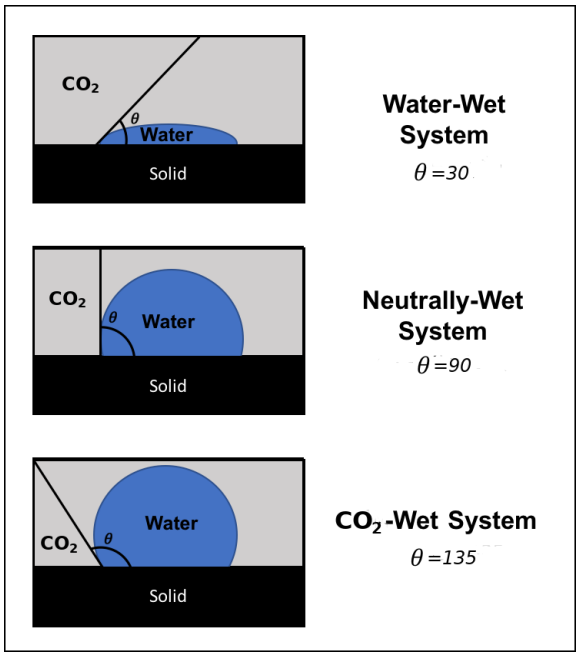


Figure 1.7: Water, CO_2 and mineral surface system illustrating different contact angles measured through the water phase.

Gradually the latter becomes denser than the adjacent reservoir fluids and falls to the bottom of the formation progressively. The mineral trapping is the result of the conversion of CO_2 into calcite because of reactions with solid minerals. CO_2 in aqueous phase produces a weak acid which reacts with mineral rocks to form bicarbonate ions characterized by different cations depending on the mineralogy of the formation. An example of reaction with potassium basic silicate is

$$3\text{K-feldspar} + 2CO_2 + 2H_2O \leftrightarrow \text{Muscovite} + 6\text{Quartz} + 2K^+ + 2HCO_3^-.$$

1.1.2.2 Other phenomena and laws

Various phenomena are involved in CO_2 injection and storage: transport phenomena (convection, diffusion and dispersion) and chemical phenomena (dissolution of CO_2 in water, acidification of the surrounding water, reactions between water and rock). There is a very strong link between these different phenomena. The injection of CO_2 and its dissolution in the water strongly influences the flow. The chemical equilibrium of the medium is modified by the dissolution of CO_2 in water, which can imply numerous chemical reactions, in particular dissolution and precipitation reactions of the rock. These reactions can modify the porosity, the permeability and thus change the characteristics of the porous medium, and therefore the flow properties.

Moreover, the storage of CO_2 in aquifers is made possible by gravity, capillary and viscous forces. These last ones are the dominant forces for the migration of the gas during the injection phase, because of the resulting pressure gradient. Then, in the post-injection phase, buoyancy and capillary forces make possible the CO_2 trapping. The drainage and imbibition-like processes during the injection and post-injection phases lead to hysteresis, i.e. a process in which capillary pressure and relative permeability curves change pathways (Figure 1.8). It is very important to the CO_2 trapping process modeling [84, 98, 129].

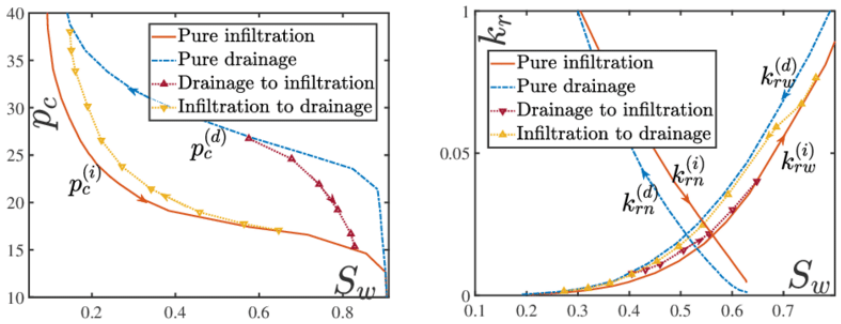


Figure 1.8: Hysteresis of capillary pressure and relative permeability curves [110].

The basic governing equations of an appropriate model are usually divided into two categories: fundamental balance laws, which express the conservation of mass and energy of the species, and empirical constitutive laws, which reflect the experimental knowledge of physicists about the physical effects under consideration. Moreover, the system can be enriched by other physical equations to predict geomechemical effects (permeability and porosity changements) and geochemical reactions, as the ones reported in the geochemical trapping paragraph.

Balance laws. The mass conservation equations connect the time rate of mass change of a species to the convection and the diffusive terms, as well as the source (sink) term, When each phase α contains only one component, the mass balance for phase α is typically

$$\partial_t(\phi\varrho_\alpha s_\alpha) + \nabla\cdot(\varrho_\alpha v_\alpha) = q_\alpha, \tag{1.1.1}$$

where $\phi \in [0,1]$ is the porosity of the medium, while $s_\alpha \in [0,1]$ is the saturation, $\varrho_\alpha > 0$ is the density, $v_\alpha \in \mathbb{R}^2$ or $\mathbb{R}^3$ is the filtration velocity vector and q_α is the source/sink term of phase α. When thermal effects are into account, one or several energy balance laws must also be envisaged. We shall neglect thermal effects in this thesis.

Darcy laws. The flow velocity of a fluid can be expressed via the Darcy law [55,112], which states that, for each phase α,

$$v_\alpha = -\lambda\frac{k_{r,\alpha}(s_\alpha, x)}{\mu_\alpha}(\nabla p_\alpha - \varrho_\alpha g), \tag{1.1.2}$$

where λ is the absolute permeability of the medium, $k_{r,\alpha} > 0$ is the relative permeability of phase α and $g \in \mathbb{R}^2$ our $\mathbb{R}^3$ is the gravity vector. The sign suggests that, in absence of gravity, the flux flows from the high pressure zone to the low pressure one. The notion of relative permeability expresses the observation that a fluid in contact with another does not move in the same way as if it were alone. It depends on the phase saturation value s_α and on the wettability and density properties of the other "contact" phases.

1.1.3 Simulation tools

In light of this brief overview of the physical and chemical processes that take place during CO_2 storage, it is easy to see the complexity of this operation. In order to avoid the risks outlined above, to predict the flow path of the injected CO_2 and to optimize the well location, extensive simulations

must be carried out before the injection phase. This is where numerical mathematics comes into action.

Normally, for modeling CO_2 storage in saline aquifers either analytical or numerical models are employed. The choice between the two depends on the purpose of the research, the given problem and the available data. Analytical models can provide information about the eligibility of a formation for storage [138], the plume migration [116], but their assumptions are oversimplified to properly describe reservoir properties and model geometry heterogeneities. Moreover the geochemical reactions occurring during the storage of CO_2 cannot be described by analytical models.

Different forms of numerical modeling techniques have been employed to modeling of the storage process such as streamline simulations, vertical equilibrium models and conventional grid-based numerical models [48,57,83,95]. Let us cite some relevant simulators. IFP Energies Nouvelles has designed the CO_2 Reservoir Environmental Simulator (COORES) research code to study CO_2 storage process from the well to the basin scale [102, 103]. COORES simulates multi-component three-phase and 3-D fluid flow in heretogeneous porous media, using structured or unstructured grids.

Another simulator project is the one developed by the University of Stuttgart: DuMu$^{\text{x}}$ [79]. It is a multi-scale, multi-physics toolbox for the simulation of flow and transport processes in porous media. It is based on the Distributed and Unified Numerics Environment (DUNE) [21, 22] which goal is to allow the use of different implementations as grids, solvers, etc. using C++ techniques. The main intention of DuMu$^{\text{x}}$ is to provide a scheme for easily and efficiently implement models for porous media flow problems. Widely used in oil and gas industry is the simulation tool ECLIPSE [125]. It includes two software packages: ECLIPSE Black Oil (E100) and ECLIPSE Compositional (E300). The first one is a fully implicit, 3D black oil simulator; the second one is a compositional simulator which have been enriched by options such as CO2STORE and GASWAT to handle CO_2 solubility in water.

1.2 Two mathematical models

The simulation of flows porous media has evolved into a mature technology, with an abundant catalog of models adapted to different needs and a wide range of numerical methods. The constant quest for a better robustness of softwares is at the root of intense research activities. In order to elaborate on the difficulties we want to address, we first need to fix ideas on the models and the schemes. In this section, we introduce two mathematical models: a "difficult" one called *two-phase system* and an "easier" one called *Richards' equation*.

We denote by Ω a bounded open set of $\mathbb{R}^d$ (with $d = 2, 3$) representing the porous medium, by $T > 0$ a finite time horizon and by

$$Q_T := \Omega \times (0, T)$$

the corresponding space-time cylinder. The boundary of the space domain is split into two parts according to

$$\partial\Omega = \Gamma^{\mathrm{D}} \cup \Gamma^{\mathrm{N}}, \qquad \Gamma^{\mathrm{D}} \cap \Gamma^{\mathrm{N}} = \varnothing,$$

and Γ^{D} having non-zero measure. One phase is called *wetting* (w) whereas the other one is called *non-wetting* (nw). The generic subscript for phase is $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$.

1.2.1 Immiscible incompressible two-phase flow in porous media

By "incompressible" it is understood that both densities ρ_α are uniform. By "immiscible" we mean that both viscosities μ_α are given constant. The two-phase system is made up of the following equations:

- The two balance laws (1.1.1) divided by ρ_α and in which $q_\alpha = 0$; more explicitly

$$\partial_t(\phi s_{\rm w}) \ + \nabla \cdot v_{\rm w} \ = 0 \qquad \text{in } Q_T, \tag{1.2.1a}$$
$$\partial_t(\phi s_{\rm nw}) + \nabla \cdot v_{\rm nw} = 0 \qquad \text{in } Q_T. \tag{1.2.1b}$$

- The two Darcy laws (1.1.2), namely,

$$v_{\rm w} + \quad \lambda \frac{k_{r,{\rm w}}(s_{\rm w}, x)}{\mu_{\rm w}} (\nabla p_{\rm w} - \varrho_{\rm w} g) \quad = 0 \qquad \text{in } Q_T, \tag{1.2.2a}$$
$$v_{\rm nw} + \lambda \frac{k_{r,{\rm nw}}(s_{\rm nw}, x)}{\mu_{\rm nw}} (\nabla p_{\rm nw} - \varrho_{\rm nw} g) = 0 \qquad \text{in } Q_T. \tag{1.2.2b}$$

- The capillary pressure-saturation relationship

$$p_{\rm nw} - p_{\rm w} =: p_{\rm c} \qquad \text{in } Q_T, \tag{1.2.3a}$$
$$s_{\rm nw} - \mathcal{S}_{\rm nw}(p_{\rm c}, x) = 0 \qquad \text{in } Q_T. \tag{1.2.3b}$$

- The volume conservation

$$s_{\rm w} + s_{\rm nw} = 1 \qquad \text{in } Q. \tag{1.2.4}$$

- The Dirichlet and Neumann boundary conditions

$$p_\alpha = p_\alpha^{\rm D} \qquad \text{on } \Gamma^{\rm D} \times (0,T), \tag{1.2.5a}$$
$$v_\alpha \cdot \nu = q_\alpha^{\rm N} \qquad \text{on } \Gamma^{\rm N} \times (0,T), \tag{1.2.5b}$$

  where $\nu \in \mathbb{R}^d$ stands for the outward unit normal vector to $\partial\Omega$.

- The initial condition

$$s_\alpha(\cdot, t=0) = s_\alpha^0 \qquad \text{in } \Omega. \tag{1.2.6}$$

The significance of various quantities has been given in §1.1.2.2. The unknowns of the system are s_α, v_α, p_α which are functions of $(x,t) \in \overline{Q}_T$.

Let us now comment on the data. The porosity ϕ and the absolute permeability λ are given functions of the space variable x. It is assumed that they both have positive lower bounds, that is,

$$\phi \geqslant \phi_{\min} > 0, \qquad \lambda \geqslant \lambda_{\min} > 0.$$

By definition (1.2.3a), the *capillary pressure* $p_{\rm c}$ is the difference between the two phase pressures. Traditionally, it is given as an empirical function of the saturation of the wetting fluid and on the type of rock present at point x. Here, for the two-phase system, we choose to take it as a function $\mathcal{S}_{\rm nw}^{-1}(\cdot, x)$ of the non-wetting saturation $s_{\rm nw}$, in order to deal with a non-decreasing capillary pressure law. The relative permeability $k_{r,\alpha}$ takes values in $[0,1]$ and is an increasing function of s_α and of the rock type present present at point x.

In the above equations, we have written $k_{r,\alpha}(s_\alpha, x)$ and $\mathcal{S}_{\rm nw}(p_{\rm c}, x)$ to stress the explicit dependency of $k_{r,\alpha}$ and $\mathcal{S}_{\rm nw}$ on location x through the rock type at this point. When several types of rock are encountered, the medium is said to be *heterogeneous*. When only one type of rock can be found, the medium is said to be *homogeneous*. In the latter case, we can simply write $k_{r,\alpha}(s_\alpha)$ and $\mathcal{S}_{\rm nw}(p_{\rm c})$. It has been proven that system (1.2.1)–(1.2.6), for suitable initial and boundary conditions, admits a solution in homogeneous domains [52, 101]. Moreover, an analysis of the limit as the constitutive relation degenerates into a maximal monotone graph is undertaken in [9]. Finally, regularity properties of the solution are investigated in [7].

1.2.2 Richards' approximation

Richards' equation is an approximation of the previously described two-phase model. It is valid in the so-called *vadose* zone (also termed *unsaturated* zone). This approximation is of interest in the field of hydrology and also in mathematics, as it allows one to work initially on a reduced model and then move on to the full model. Thanks to the assumption

$$p_{\rm nw} = 0,$$

the wetting and non-wetting phases of the two-phase system (1.2.1)–(1.2.6) can be decoupled. The resulting equations for the wetting phase are

- The volume balance law
$$\partial_t(\phi s_{\rm w}) + \nabla \cdot v_{\rm w} = 0 \qquad \text{in } Q_T. \tag{1.2.7}$$
- The Darcy law
$$v_{\rm w} + \lambda \frac{k_{r,{\rm w}}(s_{\rm w}, x)}{\mu_{\rm w}}(\nabla p_{\rm w} - \varrho_{\rm w} g) = 0 \qquad \text{in } Q_T. \tag{1.2.8}$$
- The capillary pressure-saturation relationship
$$s_{\rm w} - \mathcal{S}_{\rm w}(p_{\rm w}, x) = 0 \qquad \text{in } Q_T. \tag{1.2.9}$$
- The Dirichlet and Neumann boundary conditions
$$p_{\rm w} = p_{\rm w}^{\rm D} \qquad \text{on } \Gamma^{\rm D} \times (0,T), \tag{1.2.10a}$$
$$v_{\rm w} \cdot \nu = q_{\rm w}^{\rm N} \qquad \text{on } \Gamma^{\rm N} \times (0,T). \tag{1.2.10b}$$
- The initial condition
$$s_{\rm w}(\cdot, t=0) = s_{\rm w}^0 \qquad \text{in } \Omega. \tag{1.2.11}$$

Note that in the capillary pressure-saturation law (1.2.9), we now work with $s_{\rm w}$ and $p_{\rm w}$, which is the opposite of the capillary pressure. Since only the wetting phase is involved, we can safely omit the subscript w for notational convenience. System (1.2.7)–(1.2.11) then becomes

$$\partial_t(\phi s) + \nabla \cdot v = 0 \qquad \text{in } Q_T, \tag{1.2.12a}$$
$$v + \lambda \frac{k_r(s, x)}{\mu}(\nabla p - \varrho g) = 0 \qquad \text{in } Q_T, \tag{1.2.12b}$$
$$s - \mathcal{S}(p, x) = 0 \qquad \text{in } Q_T, \tag{1.2.12c}$$
$$p = p^{\rm D} \qquad \text{on } \Gamma^{\rm D} \times (0,T), \tag{1.2.12d}$$
$$v \cdot \nu = q^{\rm N} \qquad \text{on } \Gamma^{\rm N} \times (0,T), \tag{1.2.12e}$$
$$s(\cdot, t=0) = s^0 \qquad \text{in } \Omega. \tag{1.2.12f}$$

Elimination of v leads to the reduced system

$$\partial_t(\phi s) - \nabla \cdot \Big[\lambda \frac{k_r(s,x)}{\mu}(\nabla p - \varrho g)\Big] = 0 \qquad \text{in } Q_T, \tag{1.2.13a}$$

$$s - \mathcal{S}(p,x) = 0 \qquad \text{in } Q_T, \tag{1.2.13b}$$

$$p = p^{\mathrm{D}} \qquad \text{on } \Gamma^{\mathrm{D}} \times (0,T), \tag{1.2.13c}$$

$$-\lambda \frac{k_r(s,x)}{\mu}(\nabla p - \varrho g) \cdot \nu = q^{\mathrm{N}} \qquad \text{on } \Gamma^{\mathrm{N}} \times (0,T), \tag{1.2.13d}$$

$$s(\cdot, t=0) = s^0 \qquad \text{in } \Omega. \tag{1.2.13e}$$

The first two equations (1.2.7)–(1.2.8) can be further reduced by eliminating either the saturation to have a scalar equation in p, namely,

$$\partial_t(\phi \mathcal{S}(p,x)) - \nabla \cdot \Big[\lambda \frac{k_r(\mathcal{S}(p,x),x)}{\mu}(\nabla p - \varrho g)\Big] = 0, \tag{1.2.14}$$

or the pressure to have a scalar equation in s, namely,

$$\partial_t(\phi s) - \nabla \cdot \Big[\lambda \frac{k_r(s,x)}{\mu}(\nabla \mathcal{S}^{-1}(s,x) - \varrho g)\Big] = 0, \tag{1.2.15}$$

assuming invertibility of $\mathcal{S}(\cdot,x)$ for each fixed x.

In a homogeneous domain or inside a homogeneous subdomain, it is customary to define the *Kirchhoff transform* as

$$u = \mathscr{U}(p) := \int_{-\infty}^{p} k_r(\mathcal{S}(\pi))\,\mathrm{d}\pi. \tag{1.2.16}$$

This new quantity, also known as the *global pressure*, combines the two main nonlinearities in one and has the advantage of linearizing the gradient term, insofar as

$$\nabla u = k_r(\mathcal{S}(p))\nabla p.$$

It can be used instead of p as primary variable. Indeed, defining the new function $\mathscr{S}(u) = \mathcal{S}(p)$, we can recast the homogeneous version of (1.2.13) under the form

$$\partial_t(\phi s) - \nabla \cdot \Big[\frac{\lambda}{\mu}(\nabla u - k_r(s)\varrho g)\Big] = 0, \qquad \text{in } Q_T, \tag{1.2.17a}$$

$$s - \mathscr{S}(u) = 0, \qquad \text{in } Q_T, \tag{1.2.17b}$$

$$u = \mathscr{U}(p^{\mathrm{D}}) \qquad \text{on } \Gamma^{\mathrm{D}} \times (0,T), \tag{1.2.17c}$$

$$-\frac{\lambda}{\mu}(\nabla u - k_r(s)\varrho g) \cdot \nu = q^{\mathrm{N}} \qquad \text{on } \Gamma^{\mathrm{N}} \times (0,T), \tag{1.2.17d}$$

$$s(\cdot, t=0) = s^0 \qquad \text{in } \Omega. \tag{1.2.17e}$$

The existence and uniqueness of the solution of system (1.2.17) is studied in [8, 117].

1.2.3 Constitutive relations

In industrial applications, the most classical models used for the relative permeabilities $k_{r,\alpha}$ and the capillary pressure-saturation relation $\mathcal{S}_\alpha$ are those of Brooks-Corey [39] and van Genuchten-Mualem [133]. Both models involve various parameters such as the residual non-wetting saturation $s_{\mathrm{rn}} \in [0,1]$ and the residual wetting saturation $s_{\mathrm{rw}} \in [0,1]$ such that

$$s_{\mathrm{rn}} + s_{\mathrm{rw}} < 1. \tag{1.2.18}$$

1.2.3.1 Two-phase system

In a homogeneous domain characterized by a given pair $(s_{\mathrm{rn}}, s_{\mathrm{rw}})$, the *effective* non-wetting saturation is defined as

$$\widetilde{s}_{\mathrm{eff}} := \widetilde{s}_{\mathrm{eff}}(s_{\mathrm{nw}}) = \Pi_{[0,1]}\left(\frac{(1-s_{\mathrm{rw}})-s_{\mathrm{nw}}}{(1-s_{\mathrm{rw}})-s_{\mathrm{rn}}}\right) = \Pi_{[0,1]}\left(\frac{s_{\mathrm{w}}-s_{\mathrm{rw}}}{(1-s_{\mathrm{rn}})-s_{\mathrm{rw}}}\right), \tag{1.2.19}$$

where $\Pi_{[0,1]}(\cdot)$ denotes the projection on the interval $[0,1]$, that is,

$$\Pi_{[0,1]}(r) = \begin{cases} 0 & \text{if } r < 0, \\ r & \text{if } 0 \leqslant r \leqslant 1, \\ 1 & \text{if } r > 1. \end{cases} \tag{1.2.20}$$

Then, we have

- for the Brooks-Corey model:

$$k_{r,\mathrm{w}}(s_{\mathrm{nw}}) = \widetilde{s}_{\mathrm{eff}}^{\,3+2/n}, \tag{1.2.21a}$$

$$k_{r,\mathrm{nw}}(s_{\mathrm{nw}}) = (1-\widetilde{s}_{\mathrm{eff}})^2(1-\widetilde{s}_{\mathrm{eff}}^{\,1+2/n}), \tag{1.2.21b}$$

$$\mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}}) = \begin{cases} 1-\left[s_{\mathrm{rw}} + (1-s_{\mathrm{rn}}-s_{\mathrm{rw}})\left(\dfrac{p_{\mathrm{c}}}{p_b}\right)^{-n}\right] & \text{if } p_{\mathrm{c}} > p_b, \\ s_{\mathrm{rn}} & \text{if } p_{\mathrm{c}} \leqslant p_b; \end{cases} \tag{1.2.21c}$$

- for the van Genuchten-Mualem model:

$$k_{r,\mathrm{w}}(s_{\mathrm{nw}}) = \widetilde{s}_{\mathrm{eff}}^{\,1/2}\{1-\big[1-\widetilde{s}_{\mathrm{eff}}^{\,1/m}\big]^m\}^2, \tag{1.2.22a}$$

$$k_{r,\mathrm{nw}}(s_{\mathrm{nw}}) = (1-\widetilde{s}_{\mathrm{eff}})^{1/2}\big[1-\widetilde{s}_{\mathrm{eff}}^{\,1/m}\big]^{2m}, \tag{1.2.22b}$$

$$\mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}}) = \begin{cases} 1-\left[s_{\mathrm{rw}} + (1-s_{\mathrm{rn}}-s_{\mathrm{rw}})\left(1+\left|\dfrac{\xi p_{\mathrm{c}}}{\varrho_{\mathrm{w}} g}\right|^n\right)^{-m}\right] & \text{if } p_{\mathrm{c}} > 0, \\ s_{\mathrm{rn}} & \text{if } p_{\mathrm{c}} \leqslant 0, \end{cases} \tag{1.2.22c}$$

  with $m = 1 - 1/n$.

The purpose of the projection operator $\Pi_{[0,1]}$ in (1.2.19) is to make formulas for $k_{r,\alpha}$ valid for all $s_\alpha \in [0,1]$. In particular, we have

$$k_{r,\alpha}(s_\alpha) = \begin{cases} 0 & \text{for } s_\alpha \in [0, s_{\mathrm{r}\alpha}], \\ 1 & \text{for } s_\alpha \in [1-s_{\mathrm{r}\overline{\alpha}}, 1], \end{cases} \tag{1.2.23}$$

where $\overline{\alpha}$ stands for the complementary phase of α, i.e., $\overline{\alpha} = \mathrm{nw}$ if $\alpha = \mathrm{w}$ and $\overline{\alpha} = \mathrm{w}$ if $\alpha = \mathrm{nw}$. In the formulas for $\mathcal{S}_{\mathrm{nw}}$, the parameter $p_b > 0$ is the *entry* pressure. The function $\mathcal{S}_{\mathrm{nw}} : \mathbb{R} \to [0,1]$ can be easily seen to be nondecreasing and to satisfy

$$\mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}}) = s_{\mathrm{rn}} \qquad \text{for } p_{\mathrm{c}} \leqslant p_b, \tag{1.2.24a}$$

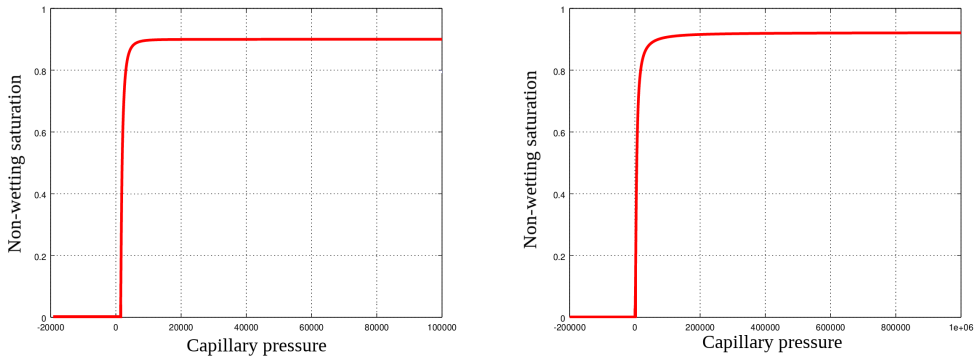
$$\mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}}) \to 1 - s_{\mathrm{rw}} \qquad \text{as } p_{\mathrm{c}} \to +\infty. \tag{1.2.24b}$$

$1 - s_{\mathrm{rn}}$	s_{rw}	$p_b\,[\mathrm{Pa}]$	n	$\lambda\,[\mathrm{m}^2]$	ϕ
1.0	0.1	$1.4708 \cdot 10^3$	3.0	10^{-11}	0.35

Table 1.1: Parameters used for the Brooks-Corey model.

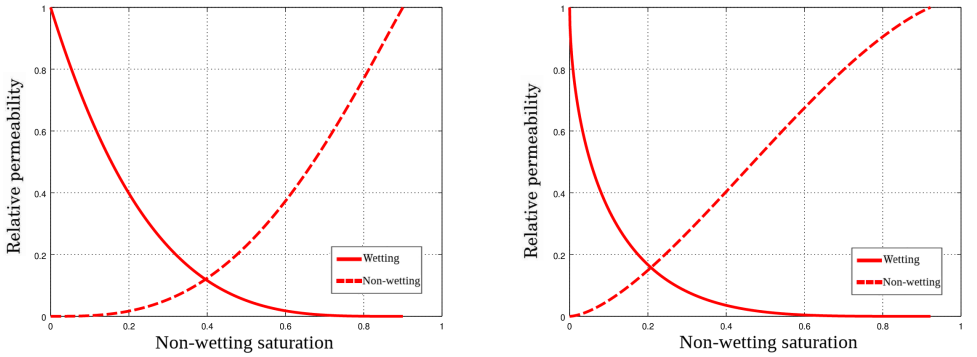
$1 - s_{\mathrm{rn}}$	s_{rw}	n	$\lambda\,[\mathrm{m}^2]$	$\xi\,[\mathrm{m}^{-1}]$	ϕ
1.0	0.0782	2.239	$6.3812 \cdot 10^{-12}$	2.8	0.3658

Table 1.2: Parameters used for the van Genuchten-Mualem model.



(a) Brooks-Corey model. (b) van Genuchten-Mualem model.

Figure 1.9: Plot of the capillary pressure curve using the Brooks-Corey model and the van Genuchten-Mualem model.



(a) Brooks-Corey model. (b) van Genuchten-Mualem model.

Figure 1.10: Plot of the wetting and non-wetting relative permeability curve using the Brooks-Corey model and the van Genuchten-Mualem model.

In Figures 1.9–1.10 we report the plot of the capillary pressure p_c and relative permeability curves $k_{r,\alpha}$ for the Brooks-Corey model and the van Genuchten-Mualem model using parameters shown in Tables 1.1–1.2 respectively.

The exponents n, m, the entry pressure p_b and the parameter ξ are also characteristic properties of the rock type. In a homogeneous domain, these are given constants. In a heterogeneous domain, the quantities

$$s_{\mathrm{rw}}(x), \quad s_{\mathrm{rn}}(x), \quad n(x), \quad m(x), \quad p_b(x), \quad \xi(x)$$

depend on the coordinate x through the local rock type. Plugging their known values into (1.2.19) and (1.2.21)–(1.2.22), we obtain

$$\widetilde{s}_{\mathrm{eff}}(s_{\mathrm{nw}}, x), \quad k_{r,\mathrm{w}}(s_{\mathrm{nw}}, x), \quad k_{r,\mathrm{nw}}(s_{\mathrm{nw}}, x), \quad \mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}}, x).$$

1.2.3.2 Richards' equation

In a homogeneous domain characterized by the residual saturations $(s_{\mathrm{rn}}, s_{\mathrm{rw}})$, the *effective* saturation is defined as

$$s_{\mathrm{eff}} := s_{\mathrm{eff}}(s) = \Pi_{[0,1]}\left(\frac{s - s_{\mathrm{rw}}}{(1 - s_{\mathrm{rn})} - s_{\mathrm{rw}}}\right), \tag{1.2.25}$$

where $\Pi_{[0,1]}$ is the projection on $[0,1]$, defined in (1.2.20). We recall that the subscript w has been dropped and that the capillary pressure-saturation is now written as $s = \mathcal{S}(p)$, where $p = -p_{\mathrm{c}}$. Then, it follows from (1.2.21)–(1.2.22) that

- for the Brooks-Corey model:

$$k_r(s) = s_{\mathrm{eff}}^{3+2/n}, \tag{1.2.26a}$$

$$\mathcal{S}(p) = \begin{cases} s_{\mathrm{rw}} + (1 - s_{\mathrm{rn}} - s_{\mathrm{rw}})\left(-\dfrac{p}{p_b}\right)^{-n} & \text{if } p < -p_b, \\ 1 - s_{\mathrm{rn}} & \text{if } p \geqslant -p_b, \end{cases} \tag{1.2.26b}$$

- for the van Genuchten-Mualem model:

$$k_r(s) = s_{\mathrm{eff}}^{1/2}\big(1 - \big[1 - s_{\mathrm{eff}}^{1/m}\big]^m\big)^2, \tag{1.2.27a}$$

$$\mathcal{S}(p) = \begin{cases} s_{\mathrm{rw}} + (1 - s_{\mathrm{rn}} - s_{\mathrm{rw}})\left[1 + \left|\dfrac{\xi p}{\varrho g}\right|^n\right]^{-m} & \text{if } p < 0, \\ 1 - s_{\mathrm{rn}} & \text{if } p \geqslant 0, \end{cases} \tag{1.2.27b}$$

  with $m = 1 - 1/n$.

Thanks to the projection operator $\Pi_{[0,1]}$, the formulas for k_r are valid for all $s \in [0,1]$. In particular, we have

$$k_r(s) = \begin{cases} 0 & \text{for } s \in [0, s_{\mathrm{rw}}], \\ 1 & \text{for } s \in [1 - s_{\mathrm{rn}}, 1]. \end{cases} \tag{1.2.28}$$

The function $\mathcal{S} : \mathbb{R} \to [0,1]$ is nondecreasing and satisfies

$$\mathcal{S}(p) = 1 - s_{\mathrm{rn}} \qquad \text{for } p \geqslant -p_{\mathrm{b}}, \tag{1.2.29a}$$

$$\mathcal{S}(p) \to s_{\mathrm{rw}} \qquad \text{as } p \to -\infty. \tag{1.2.29b}$$

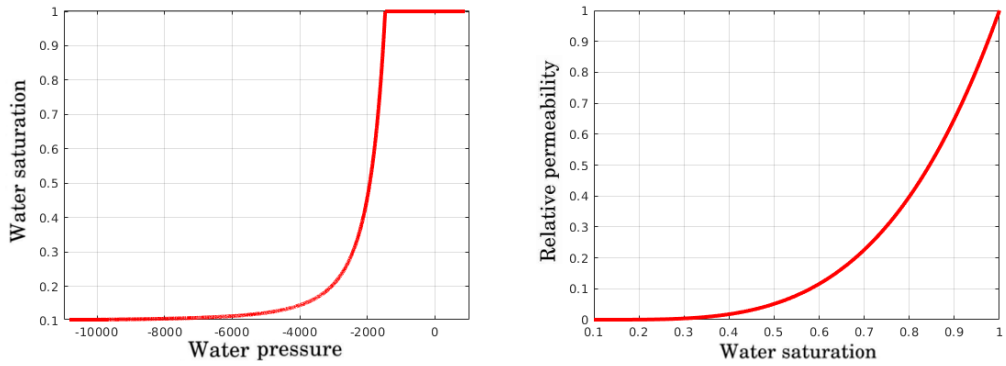


Figure 1.11: Water pressure and relative permeability curves for the Brooks-Corey model using parameters reported in Table 1.1.

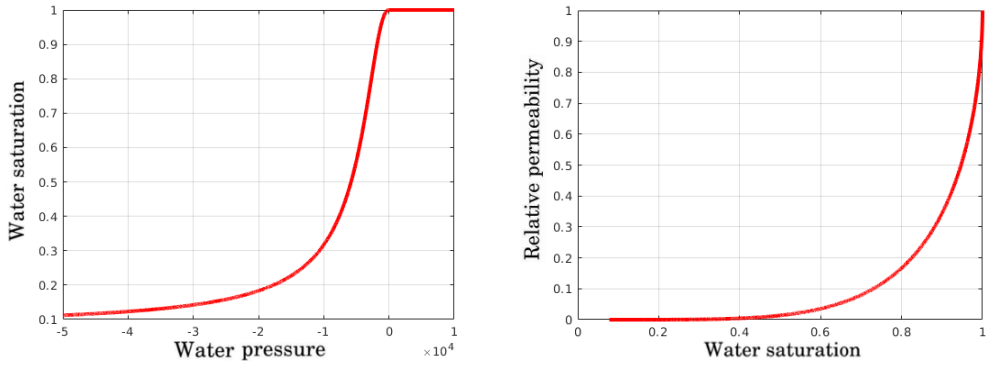


Figure 1.12: Water pressure and relative permeability curves for the van Genuchten-Mualem model using parameters reported in Table 1.2.

In Figures 1.11–1.12, we report the plot of the water pressure $p = -p_c$ and relative permeability curves k_r for the Brooks-Corey model and the van Genuchten-Mualem model using parameters shown in Tables 1.1–1.2 respectively.

In a heterogeneous domain, the quantities

$$s_{\mathrm{rw}}(x), \quad s_{\mathrm{rn}}(x), \quad n(x), \quad m(x), \quad p_b(x), \quad \xi(x)$$

depend on the coordinate x through the local rock type. Plugging their known values into (1.2.25) and (1.2.26)–(1.2.27), we obtain

$$s_{\mathrm{eff}}(s,x), \quad k_r(s,x), \quad \mathcal{S}(p,x).$$

Going back to a homogeneous domain or subdomain, let us derive the relationship $s = \mathscr{S}(u)$ when the Kirchhoff transform is applied along the lines of (1.2.16)–(1.2.17). Because the latter cannot be analytically computed for the van Genuchten-Mualem model, we just do it for the Brooks-Corey model. By straightforward calculations, we obtain

$$\mathscr{S}(u) = \begin{cases} s_{\mathrm{rw}} + (1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}) \left(\dfrac{u}{u_b} \right)^{n/(3n+1)} & \text{if } u \leqslant u_b, \\ 1 - s_{\mathrm{rn}} & \text{if } u > u_b, \end{cases} \tag{1.2.30}$$

with $u_b = -p_b/(3n+1)$. The behavior of such a function is illustrated in Figure 1.13.

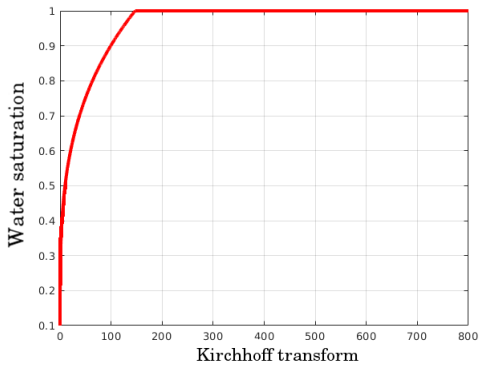


Figure 1.13: Profile of the wetting saturation-global pressure graph for Brooks-Corey model using parameters reported in Table 1.1.

1.3 Towards more robust and accurate numerical approximations

1.3.1 Two critical difficulties

We are going to explain the difficulties that arise in the numerical resolution of the Richards equation (1.2.13) in the saturation-pression formulation. Those for the Richards equation (1.2.17) with the Kirchhoff transform and for the two-phase system (1.2.1)–(1.2.6) are similar in essence.

Richards' equation is a nonlinear, degenerate elliptic-parabolic partial differential equation. Loss of ellipticity stems from the fact that k_r is allowed to vanish, by virtue of (1.2.28) and as highlighted in the left panel of Figure 1.14. This is aggravated by the low regularity of the constitutive laws k_r (at $s_{\rm rw}$ and $1 - s_{\rm rn}$) and $\mathcal{S}$ (at $p = -p_b$), as underlined in Figure 1.15. However, the worse is yet to come.

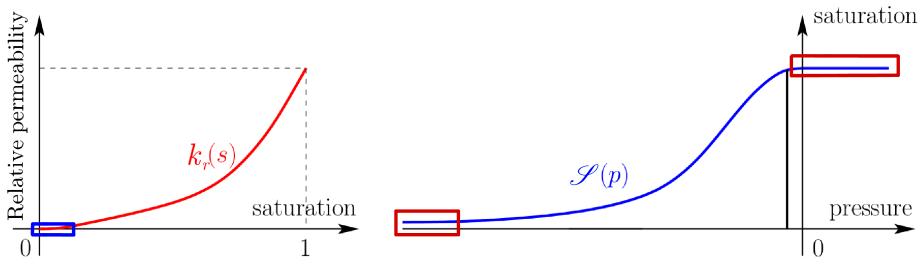
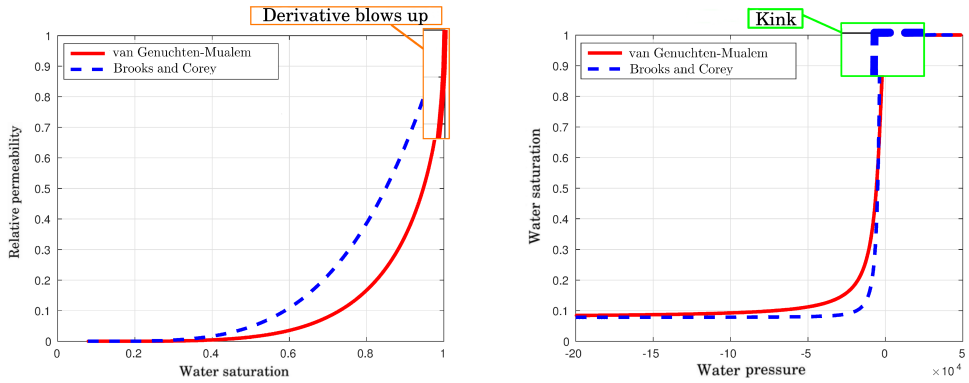


Figure 1.14: Relative permeability law k_r (left) and saturation-pression relationship $\mathcal{S}(p)$ (right).

1.3.1.1 Lack of robustness due to stiffness of closure laws

Even in the regions where k_r and $\mathcal{S}$ are smooth, other delicate issues contribute to make its numerical resolution challenging. On the one hand, when the saturation s reaches value 1, the function $\mathcal{S}$ is no longer invertible and its derivative is equal to zero, which causes the Richards equation to degenerate from parabolic to elliptic. At the practical level, this implies that the saturation s cannot



(a) Blow-up in $k_r'(s)$ for van Genuchten-Mualem. (b) Kink in $\mathcal{S}(p)$ for Brooks-Corey.

Figure 1.15: Close-up on the severe stiffness of the constitutive laws $k_r(s)$ and $\mathcal{S}(p)$.

be taken as a primary unknown for this regime of solution. In the same vein, it is not recommended to choose s as the primary variable in dry zones ($s \ll 1$) where $\mathcal{S}$ is very flat, as illustrated in the right panels of Figures 1.14–1.15. Indeed, inversion in p becomes ill-conditioned in the sense that $(\mathcal{S}^{-1})'(s) \to \infty$.

On the other hand, assuming that the pressure p is taken as the primary unknown, there is no guarantee the algebraic system of discretized equations can be solved in an efficient way, say, by the Newton method. To expand on this matter, let us consider a general discrete approximation space $\mathcal{X}_h$ for the pressure p. Using a backward Euler method in time, discretization methods in space usually amount to finding, at each time step n, a $p_h^n \in \mathcal{X}_h$ such that

$$\left\langle \phi(\mathcal{S}(p_h^n) - \mathcal{S}(p_h^{n-1})), v_h \right\rangle + \Delta t^n \left\langle \frac{\lambda k_r(\mathcal{S}(p_h^n))}{\mu} (\nabla p_h^n - \varrho g), \nabla v_h \right\rangle + \ldots = 0 \tag{1.3.1}$$

for all $v_h \in \mathcal{X}_h$. In (1.3.1), $\langle \cdot, \cdot \rangle$ denotes a discrete scalar product related to a given scheme, $\Delta t^n = t^n - t^{n-1}$ and the dots may include other terms like boundary conditions (or even sources) we do not detail. In the sequel, we will assume that these terms are absent without loss of generality.

The most classical way to solve equation (1.3.1) consists in using Newton's method. Starting from $p_h^{n,0} = p_h^{n-1}$, at each iteration $k \geqslant 0$ find

$$p_h^{n,k+1} = p_h^{n,k} + \delta_h^{n,k}$$

such that the increment $\delta_h^{n,k}$ solves the linearized system

$$\begin{aligned} &\left\langle \phi \mathcal{S}(p_h^{n,k}), v_h \right\rangle + \left\langle \phi \mathcal{S}'(p_h^{n,k}) \delta_h^{n,k}, v_h \right\rangle + \Delta t^n \left\langle \lambda \frac{k_r(\mathcal{S}(p_h^{n,k}))}{\mu} (\nabla p_h^{n,k+1} - \varrho g), \nabla v_h \right\rangle \\ &\qquad + \Delta t^n \left\langle \lambda \frac{k_r'(\mathcal{S}(p_h^{n,k}))}{\mu} \mathcal{S}'(p_h^{n,k}) \delta_h^{n,k} (\nabla p_h^{n,k} - \varrho g), \nabla v_h \right\rangle \\ &\qquad = \left\langle \phi \mathcal{S}(p_h^{n-1}), v_h \right\rangle. \end{aligned} \tag{1.3.2}$$

Under favorable conditions, the Newton method converges quadratically provided that the starting point is close enough to the exact solution. Here, we are very far from this ideal situation. Troubles occur with the linear system (1.3.2) itself, since the matrix on the left-hand side contains the

derivatives $k_r'(\mathcal{S}(p_h^{n,k}))$ and $\mathcal{S}'(p_h^{n,k})$, each of which may become very large or even blow up, as shown in Figure 1.15. This testifies to a highly ill-conditioned problem, which in practice leads to a failure of convergence for the Newton method.

1.3.1.2 Lack of accuracy due to strong heterogeneities

Heterogeneous domains, as suggests the adjective, are characterized by the presence of different lithologies presenting, consequently, piecewise-uniform petrophysical properties. Each lithology i gives rise to a pair $(k_{r,i}, \mathcal{S}_i)$ of stiff closure laws [23, §5.1]. A strong contrast in the parameters of these laws across an interface gives rise to an additional difficulty in the numerical resolution of the model, and appears therefore as a challenge to be taken up.

Indeed, it is observed (§3–§4) that not only the robustness of the nonlinear iterative solver is worsened, but also the orders of convergence of standard numerical methods become extremely low with respect to the mesh size (e.g., 0.3 instead of 1, as examplified in Figure 1.16 for the test case of §4.4.1.1). The heart of the matter is that the transmission conditions, especially pressure continuity, are only recovered asymptotically. As a consequence, a naive scheme without any specific treatment for heterogeneities would suffer from a lack of accuracy in the predicted results.

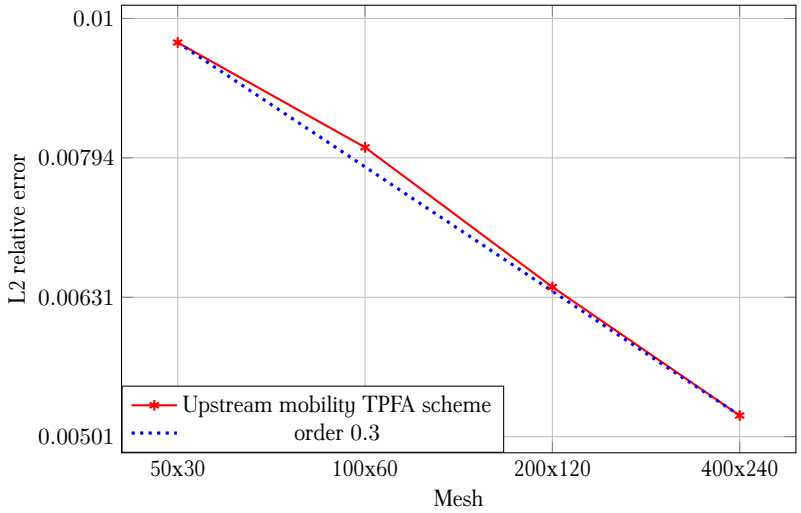


Figure 1.16: $L^2(Q_T)$ relative error in saturation for a filling test case using the Brooks-Corey model for a domain with two rock types: sand and clay.

Discontinuities in the capillary pressure function across the interface between different geological layers may yield discontinuous saturations, since the phase pressures (thus the capillary pressure) have to remain continuous provided both phases are present on both sides of the interface. The effects of such discontinuities have been discussed for instance in [25, 113, 132], while mathematical contributions concerning the analysis of this phenomenon have been proposed in [27, 42, 46]. These discontinuities are at the basis of the *capillary barrier* phenomenon illustrated in Figure 1.17. The capillary barrier effect plays a chief role for flows in porous media and in fractured ones in particular. This is why an improved accuracy in its computation is a major issue for engineers.

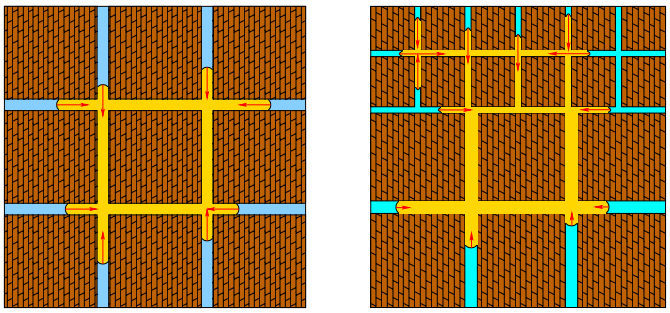


Figure 1.17: Within a homogeneous ideal porous medium with constant pore size, the resulting capillary force on a connected component of non-wetting phase vanishes (left). This is no longer true if the non-wetting phase straddles the interface between two idealized porous media (right): capillarity generates a force orthogonal to the interface. Source: [13]

1.3.2 Classical approaches

Let us review some of the numerous attempts that have been made in the literature to address the two above difficulties. Again, for the sake of simplicity, the presentation is made in an informal way for the Richards equation.

1.3.2.1 For the robustness of nonlinear iterative solvers

Alternatives to Newton's method. In place of the original Newton iterate (1.3.2), many variants have been advocated to ensure a greater robustness, often at the expense of the rate of convergence.

The modified Picard method, proposed by Celia et al. [49], deliberately omits the derivative k_r' of the relative permeability. In other words, the linear system to be solved at each iteration becomes

$$\begin{aligned}\big\langle \phi \mathcal{S}(p_h^{n,k}), v_h \big\rangle + \big\langle \phi \mathcal{S}'(p_h^{n,k}) \delta_h^{n,k}, v_h \big\rangle \\ + \Delta t^n \Big\langle \lambda \frac{k_r(\mathcal{S}(p_h^{n,k}))}{\mu} (\nabla p_h^{n,k+1} - \varrho g), \nabla v_h \Big\rangle = \big\langle \phi \mathcal{S}(p_h^{n-1}), v_h \big\rangle. \end{aligned} \tag{1.3.3}$$

By removing k_r' to avoid infinity in the Jacobian matrix, we end up with a *quasi-Newton* method, the convergence of which is notoriously slower. But at least iterations do not stop prematurely and their convergence appears to be less dependent on the mesh size than for Newton's case [107].

The original Newton method and the modified Picard method can be hybridized, as suggested by Lehmann and Ackerer [104]: its basic version consists in performing a few iterations with the modified Picard method and then in switching to Newton's scheme, once we have

$$\|\delta_h^{n,i} - \delta_h^{n,i-1}\| \leqslant \delta_a + \delta_r \|\delta_h^{n,i}\|.$$

Since δ_a and δ_r cannot always be prescribed easily beforehand, a fixed number of Picard iterations may rather be performed at each time-step. Note that in case the time derivative is not linearized in (1.3.3), the corresponding nonlinear system also gives a conservative approximation of p_h.

Casulli and Zanolli [47] put forward the more sophisticated idea of nested Newton iterations, assuming that the saturation-pressure relationship can be decomposed into the form $\mathcal{S} = \mathcal{S}_1 - \mathcal{S}_2$, where the derivatives of $\mathcal{S}_1$ and $\mathcal{S}_2$ are both nondecreasing. The derivative k_r' is neglected, as in

the modified Picard method. Thus, if k designates the current Picard iterate, ℓ and m the inner and outer Newton iterates, the linearized procedure is now

$$\begin{aligned}-\big\langle\phi\mathcal{S}_2(p_h^{n,k,\ell}),v_h\big\rangle-\big\langle\phi\mathcal{S}_2'(p_h^{n,k,\ell})\delta_{2,h}^{n,k,\ell,m},v_h\big\rangle\\ +\big\langle\phi\mathcal{S}_1(p_h^{n,k,\ell+1,m}),v_h\big\rangle+\big\langle\phi\mathcal{S}_1'(p_h^{n,k,\ell+1,m})\delta_{1,h}^{n,k,\ell,m},v_h\big\rangle\\ +\Delta t^n\left\langle\lambda\frac{k_r(\mathcal{S}(p_h^{n,k}))}{\mu}(\nabla p_h^{n,k,\ell,m}-\varrho g),\nabla v_h\right\rangle=\big\langle\phi\mathcal{S}(p_h^{n-1}),v_h\big\rangle\end{aligned}\tag{1.3.4}$$

with

$$\delta_{1,h}^{n,k,\ell,m}=p_h^{n,k,\ell+1,m+1}-p_h^{n,k,\ell+1,m},\qquad\delta_{2,h}^{n,k,\ell,m}=p_h^{n,k,\ell+1,m+1}-p_h^{n,k,\ell}.$$

Both Newton loops stop once the nonlinearities related to $\mathcal{S}_1$ and $\mathcal{S}_2$ have been solved.

Another quasi-Newton method is the *L-scheme*, in reference to the Lipschitz constant that is chosen as an abrupt substitute for the exact derivative of the accumulation term [118,128]. Replacing $S'(p_h^{n,k})$ in the modified Picard iterate (1.3.3) by $L=\sup_p|\mathcal{S}'(p)|$, this scheme writes

$$\begin{aligned}\big\langle\phi\mathcal{S}(p_h^{n,k}),v_h\big\rangle+\big\langle\phi L\delta_h^{n,k},v_h\big\rangle\\ +\Delta t^n\left\langle\frac{\lambda k_r(\mathcal{S}(p_h^{n,k}))}{\mu}(\nabla p_h^{n,k+1}-\varrho g),\nabla v_h\right\rangle=\big\langle\phi\mathcal{S}(p_h^{n-1}),v_h\big\rangle.\end{aligned}\tag{1.3.5}$$

List and Radu [107] proved that this scheme is unconditionnally linearly convergent with a rate depending on L, the value of the time step and the mobility function $\lambda k_r/\mu$ but not on the mesh size. In the same way as previously mentioned, this scheme can also be combined with Newton's one in order to accelerate convergence.

Preconditioned Newton's method. Another philosophy to gain in robustness is to keep Newton's method as the linearization scheme while judiciously preconditioning the system to be solved. Let us recall that for a linear system

$$\mathbf{A}w=b,\tag{1.3.6}$$

a preconditioner is an invertible matrix $\mathbf{P}$ such that either product $\mathbf{P}^{-1}\mathbf{A}$ or $\mathbf{A}\mathbf{P}^{-1}$ has a smaller condition number. We also recall that the latter is involved in the upper bound on the error on the solution of (1.3.6) generated by a small pertubation on the data. Rather than working with the original system (1.3.6), we can consider

- the left-preconditioned system
$$(\mathbf{P}^{-1}\mathbf{A})w=\mathbf{P}^{-1}b,\tag{1.3.7}$$
which alters the matrix and the right-hand side, but not the unknown;
- the right-preconditioned system
$$(\mathbf{A}\mathbf{P}^{-1})z=b,\qquad w=\mathbf{P}^{-1}z,\tag{1.3.8}$$
which alters the matrix and the unknown, but not the right-hand side.

Ideally, a good preconditioner $\mathbf{P}$ should be close to the original matrix $\mathbf{A}$ while being easy and inexpensive to invert.

For a nonlinear "square" system

$$\mathcal{F}(w)=0,\tag{1.3.9}$$

we can analogously define a preconditioner to be an invertible function $\mathcal{G}$ such that the Jacobian matrix of either $\mathcal{G}^{-1} \circ \mathcal{F}$ or $\mathcal{F} \circ \mathcal{G}^{-1}$ has a better conditioning than that of $\mathcal{F}$. This opens the way to considering

- the left-preconditioned system

$$(\mathcal{G}^{-1} \circ \mathcal{F})(w) = \mathcal{G}^{-1}(0); \tag{1.3.10}$$

  Brenner [28] implemented left-preconditioning for Richards' equation.

- the right-preconditioned system

$$(\mathcal{F} \circ \mathcal{G}^{-1})(z) = 0, \qquad w = \mathcal{G}^{-1}(z); \tag{1.3.11}$$

  the parametrization technique of Brenner and Cancès [29] for Richards' equation in the Kirchhoff transform formulation is in fact equivalent to right-preconditioning.

1.3.2.2 For the accuracy in heterogeneous domains

Regarding the numerical approximation in heterogeneous domains, a conforming P_1 finite element method has been proposed in [69]. Mixed Hybrid Finite Element and discontinuous Galerkin schemes have respectively been proposed in [67,90,111]. There are more contributions in the realm of Finite Volumes. In [65], a two-point flux approximation (TPFA) scheme for a simplified model was studied. It was extended in [41] to the case of multivalued capillary pressure graphs in the one-dimensional setting, and then in [30] to the multi-dimensional setting. The generalization to more general schemes allowing for anisotropy and general grid was carried out in [74] in the general framework of Gradient Schemes [62]. Similar approaches have been applied to hybrid-dimensional models for flows in porous media with fracture networks [34,64].

The convergence of the nonlinear solvers are discussed in [37, 88, 94, 135]. In particular, the papers [37, 88] illustrate the better robustness in terms of nonlinear solvers of the so-called Hybrid Upwinding Method —where two different upwindings are selected for the countercurrent contributions (buoyancy and capillary diffusion) and the global convection driving both phases in the same direction — with respect to the more classical Phase Potential Upwinding approach where each phase is upwinded with respect to its own velocity.

Finally, let us mention the contributions [10,11,13] on the numerical simulation of the vanishing capillarity limit of the equations, and the contributions [45,113,136] for vertically integrated reduced models accounting for capillary trapping.

1.3.3 Contributions and outline of this thesis

We sketch out the objectives as well as the methodological approach adopted. As before, the discussion is set for the Richards equation, but the ideas are also relevant for the two-phase system.

1.3.3.1 Robust Newton solver based on parametrization technique

As was pointed out in §1.3.1.1, the saturation s is not a good choice of primary variable in the dry soils ($s \ll 1$) or when the pressure p exceeds its entry value ($p > -p_b$). Nevertheless, s turns out to be a legitimate and convenient primary variable in the sharp transition region from s_{rw} to $1 - s_{\mathrm{rn}}$, where $\mathcal{S}'(p)$ grows larger and larger (Figure 1.15b). The fact that p should be taken as

a natural primary variable somewhere and s should play this role somewhere else has motivated practitioners to devise schemes based on variable switch between s and p, see [58,81].

The starting point of our approach is the work of Brenner and Cancès [29], who reformulated the variable switch as a parametrization of the graph $\{p, \mathcal{S}(p)\}$. An admissible parametrization is given by two nondecreasing functions

$$\mathfrak{s} : I \to [s_{\mathrm{rw}}, 1 - s_{\mathrm{rn}}], \qquad \mathfrak{p} : I \to \mathbb{R}, \tag{1.3.12}$$

where $I \in \mathbb{R}$ is some appropriate interval, such that

$$\mathfrak{s}(\tau) = \mathcal{S}(\mathfrak{p}(\tau)), \qquad 0 < \mathfrak{s}'(\tau) + \mathfrak{p}'(\tau) < \infty, \tag{1.3.13}$$

for all $\tau \in I \subset \mathbb{R}$. The last condition ensures that we cannot have $\mathfrak{s}'(\tau) = \mathfrak{p}'(\tau) = 0$ (which would make the Jacobian matrix singular) and that both derivatives $(\mathfrak{s}'(\tau), \mathfrak{p}'(\tau))$ remain bounded (which avoids blow-up). This original abstract viewpoint not only makes it easier to implement the switching, but also paves the way to a whole new family of right preconditioners, in the sense of §1.3.2.1.

Our goal is to push further this idea, which in [29] was applied only to the Richards equation in the Kirchhoff transform formulation. The Kirchhoff transform is a convenient trick to get rid of the difficulty related to the blow-up of $k_r'(s)$. Unfortunately, it is often not possible to calculate it analytically. Nor is it always the best physical choice. This is the reason why we shall focus more heavily on the pressure-saturation formulation. In this context and because of the possible blow-up of $k_r'(s)$, it is not at all obvious that the parametrization technique will continue to work properly. In fact, as will be seen in §2, other ingredients will have to be developed.

1.3.3.2 Accurate transmission conditions for heterogeneous domains

As mentioned in §1.3.1.2, we believe that the root of the difficulties arising in heterogeneous domains lies in the violation of the transmission conditions between two lithologies occupying two subdomains Ω_i and Ω_j. Let i and j be the rock types on the two sides of the interface. The transmission conditions across this interface read

$$v_i \cdot \nu_i + v_j \cdot \nu_j = 0, \tag{1.3.14a}$$

$$p_i - p_j = 0, \tag{1.3.14b}$$

where p_i and v_i are the trace at the considered interface of the pressure $p|_{\Omega_i}$ and the flux $v|_{\Omega_i}$ respectively. In other words, they express the equality of pressure and of normal velocity.

Our approach is to enforce (1.3.14) either exactly or at least with an improved accuracy in the numerical resolution of the Richards equation, that can be written in each subdomain i as

$$\phi_i \, \partial_t s + \nabla \, v = 0, \tag{1.3.15a}$$

$$v + \lambda_i \, \eta_i(s) \, \nabla(p - \varrho g \cdot x) = 0, \tag{1.3.15b}$$

$$s - \mathcal{S}_i(p) = 0. \tag{1.3.15c}$$

To this end, after a proof of convergence for the standard finite volume scheme in the heterogeneous setting (§3), we shall work out 4 methods (§4) aimed at achieving (1.3.14) in a more satisfactory way. This class of special methods for the interface will then be applied to the Richards equation (§4) and extended to the two-phase system (§5).

1.3.3.3 Organization of the manuscript

The rest of this dissertation roughly follows an order of increasing difficulty.

We start in chapter §2 with the easier model in the easier configuration, namely, the Richards equation in homogeneous media. The simplicity brought by the homogeneity of the domain allows us to focus on the difficulty associated with stiff closure laws. We advocate some necessary adaptations of the Brenner-Cancès parametrization technique for the pressure-saturation formulation. Finally, we make the required comparisons between the two formulations and between several choices of primary variable.

Heterogeneity begins to appear in Chapter §3, also for the Richards equation. The purpose of this chapter is purely theoretical. It is about proving the convergence of the standard TPFA (Two-Point Flux Approximation) finite volume scheme in heterogeneous domain, without any specific treatment of the interfaces between different rocks. The text of this chapter is replicated from the accepted article [20].

The numerical aspects of the Richards equation in heterogeneous media are the subject of chapter §4, where we numerically demonstrate that the basic TPFA scheme actually gives rise to a very low order of convergence. To remedy this shortcoming, we derive four different strategies whose common point is the attempt to satisfy the transmission conditions at the interfaces, either exactly or approximately. A comparison of these methods, labeled A, B, C, and D, is provided along with a thorough discussion. The text of this chapter is a reproduction of the submitted paper [19].

Equipped with appropriate algorithmic tools, we can finally address in chapter §5 the more difficult model in the more difficult configuration, namely, the immiscible incompressible two-phase system in a heterogeneous domain. The parametrization technique and the four methods for transmission conditions at interface are straightforwardly extended to the new model. Here, the focus is on numerical simulations, with three validation test cases inspired by realistic operating conditions, analyzed and commented at length.

Chapter 2

Finite volume approximation of Richards' equation in homogeneous domains

Let us start our journey by the easier case of the easier model, namely, the Richards equation (1.2.13) or (1.2.17) in a homogeneous domain. Homogeneity is a simplifying assumption that enables us to focus on the difficulty brought about by the stiffness of the closure laws (relative permeability k_r and water pressure-saturation $\mathcal{S}$).

We first detail the selected numerical scheme (§2.1) together with the parametrization technique (§2.2) and we describe its resolution via the Newton method (§2.3). Then, different numerical tests are presented (§2.4) to demonstrate the robustness of the approach by means of comparative results obtained with pressure/Kirchhoff transform formulation and with/without the parametrization technique.

2.1 Finite volume scheme for the Richards equation

2.1.1 State of the art

There is a great variety of families of numerical methods for Richards' equation: finite differences [49], finite volumes [73] or finite elements [66]. Over the past years, a large number of schemes have been proposed to better approximate diffusion terms. These new methods aim at: (i) using meshes with fewer constraints on the elements' shape; (ii) taking better into account anisotropies and heterogeneities in material properties; (iii) giving consistent flux approximations in both previous cases. They include high-order methods like, for instance, discontinuous Galerkin methods [123] and low-order methods such as finite-volume methods where the flux approximation uses an extended stencil or additional unknowns.

In this work, we consider the simplest scheme within this finite volume family, that is, the *Two-Point Flux Approximation* (TPFA). On a given face, this approximation only uses the two unknowns of the cells located on each other side. Its name follows from this reduced stencil. Unfortunately, this approximation does not provide consistent fluxes for general meshes and diffusion tensors. To make this approximation valid we need and isotropic media (diagonal diffusion tensor) and an admissible mesh in the sense of Definition 2.1.1. The TPFA is however often used in practice despite these flaws since the maximum principle is preserved.

To ensure consistent approximations, other methods could be used which do not necessarily maintain the positivity of the solutions nor the maximum principle. Some of them are presented in

the review [60]: the Multi Point Flux Approximations (MPFA) [100] which make use of an extended stencil, the Hybrid Mimetic Mixed Methods which include the Mimetic Finite Differences and the Hybrid and Mixed Methods [36, 63] and introduce additional face unknowns, or Discrete Duality Finite Volumes [17]. The Vertex Approximate Gradient (VAG) scheme, which uses both cell and nodal unknowns, has also been extensively studied for two-phase flows and Richards problems [74] and for fractured media [33].

We also point out that non-linear finite volume schemes, with TPFA [60, 73] and MPFA [1, 2] approximations, can also be envisaged. Their non-linearity comes from the scheme coefficients, the so-called transmissivities, which non-linearly depend on the discrete local unknowns. Despite the fact that these schemes require the use of a few Newton or Picard iterations to compute approximations of the solutions to continuous linear problems, it has been shown that they lead to positive solutions or solutions respecting the maximum principle [126].

Another important issue related to the spatial discretization is the treatment of convection terms. Classically, an upstream choice is made to compute the mobility functions according to the sign of the whole flux. Recent works have shown that specific upwindings knwon as *hybrid upwinding* and designed for the gravity and capillarity terms (and for the total velocity term for two-phase flows [32]) can improve the convergence of Newton's algorithm [88].

2.1.2 Mesh and time-steps

Definition 2.1.1. *An* admissible mesh of Ω *(see Figure 2.1) is a triplet $(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}})$ such that the following conditions are fulfilled:*

(i) Each control volume (or cell) $K \in \mathscr{T}$ is non-empty, open, polyhedral and convex, with positive d-dimensional Lebesgue measure $m_K > 0$. We assume that

$$K \cap L = \varnothing \quad \text{if } K, L \in \mathscr{T} \text{ with } K \neq L, \qquad \text{while} \quad \bigcup_{K\in\mathscr{T}} \overline{K} = \overline{\Omega}.$$

(ii) Each face $\sigma \in \mathscr{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff measure $m^{d-1}(\sigma) = m_\sigma > 0$. We assume that $m^{d-1}(\sigma \cap \sigma') = 0$ for $\sigma, \sigma' \in \mathscr{E}$ unless $\sigma' = \sigma$. For all $K \in \mathscr{T}$, we assume that there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \bigcup_{\sigma\in\mathscr{E}_K} \sigma$. Moreover, we suppose that $\bigcup_{K\in\mathscr{T}} \mathscr{E}_K = \mathscr{E}$. Given two distinct control volumes $K, L \in \mathscr{T}$, the intersection $\overline{K} \cap \overline{L}$ either reduces to a single face $\sigma \in \mathscr{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

(iii) The cell-centers $(x_K)_{K\in\mathscr{T}}$ are pairwise distinct with $x_K \in K$, and are such that, if $K, L \in \mathscr{T}$ share a face $K|L$, then the vector $x_L - x_K$ is orthogonal to $K|L$.

(iv) For the boundary faces $\sigma \subset \partial\Omega$, we assume that either $\sigma \subset \Gamma^{\mathrm{D}}$ or $\sigma \subset \overline{\Gamma}^{\mathrm{N}}$. For $\sigma \subset \partial\Omega$ with $\sigma \in \mathscr{E}_K$ for some $K \in \mathscr{T}$, we assume additionally that there exists $x_\sigma \in \sigma$ such that $x_\sigma - x_K$ is orthogonal to σ.

The set of edges $\mathscr{E}$ is then subdivided into: the set of internal faces shared by two cells $\mathscr{E}_{int} = \{\sigma = K|L \in \mathscr{E} \mid K, L \in \mathscr{T}\}$, and the set of boundary edges $\mathscr{E}_{\mathrm{ext}} = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \partial Omega\}$. This last set includes the set of Dirichlet faces $\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \Gamma^{\mathrm{D}}\}$ and the set of Neumann faces $\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}} = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \Gamma^{\mathrm{N}}\}$. We also introduce the local set $\mathscr{E}_K = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \partial K\}$ containing all the faces surrounding a cell K. To each face $\sigma \in \mathscr{E}$ we associate a distance d_σ defined by

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathscr{E}_{\mathrm{int}}, \\ d_{K,\sigma} & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}) \end{cases} \tag{2.1.1}$$

where, for all pair (K,σ) such that $\sigma \in \mathscr{E}_K$, $d_{K,\sigma} = |x_K - x_\sigma|$, with x_K the cell center and x_σ the face center, which is chosen as the intersection of $[x_K, x_L]$ with σ. Moreover, for each cell K, we denote by m_K its Lebesgue measure, and by m_σ the measure of a face σ.

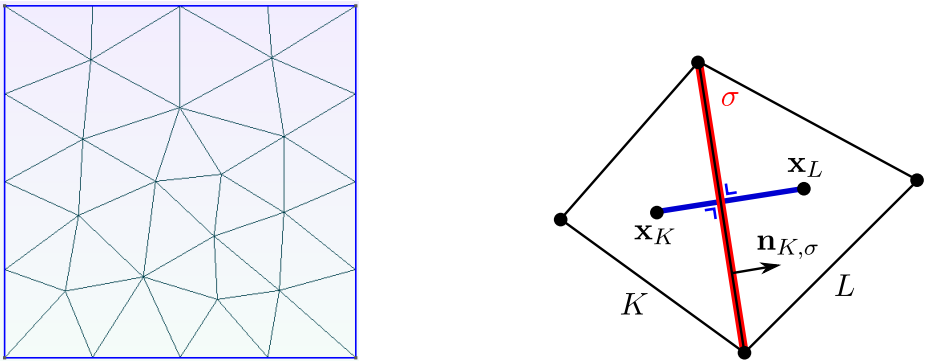


Figure 2.1: Example of admissible mesh and detail of two cells K, L sharing a face $\sigma \in \mathscr{E}_{\rm int}$.

The time discretization is given by a vector of values $(t^N)_{0\leqslant 1\leqslant N}$ with $0 = t^0 < t^1 < \cdots < t^N = T$, and we denote by $t^n - t^{n-1} = \Delta t^n$, $1 \leqslant n \leqslant N$, the time steps. At initial time $t = 0$, s^0 is discretized into

$$s_K^0 = \frac{1}{m_K}\int_K s^0 \,\mathrm{d}x.$$

We finally introduce the following notation: considering a generic variable w we define the mirror value $w_{K\sigma}^n$ of w_K^n across σ by

$$w_{K\sigma}^n = \begin{cases} w_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ w_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \\ w_\sigma^n = \dfrac{1}{\Delta t^n m_\sigma}\displaystyle\int_{t^{n-1}}^{t^n}\mathrm{d}t\int_\sigma u^{\rm D}\mathrm{d}\gamma & \text{if } \sigma \in \mathcal{E}_{\rm ext}^{\rm D}. \end{cases} \tag{2.1.2}$$

2.1.3 Implicit TPFA discretization of the model

Assuming that the mesh meets the requirement of Definition 2.1.1, let us write down the discretized equations for the Kirchhoff transform-saturation formulation first, after which those for the pressure-saturation formulation will follow in a similar fashion.

Kirchhoff transform-saturation formulation. We first put system (1.2.17) under the form

$$\partial_t(\phi s) + \nabla\cdot v = 0 \qquad \text{in } \Omega\times(0,T), \tag{2.1.3a}$$
$$v + \frac{\lambda}{\mu}(\nabla u - k_r(s)\varrho g) = 0, \tag{2.1.3b}$$
$$s - \mathscr{S}(u) = 0, \tag{2.1.3c}$$
$$u = u^{\rm D} \qquad \text{on } \Gamma^{\rm D}\times(0,T), \tag{2.1.3d}$$
$$v\cdot\nu = q^{\rm N} \qquad \text{on } \Gamma^{\rm N}\times(0,T). \tag{2.1.3e}$$

It is convenient to introduce the gravity potential

$$\psi = -\varrho g\cdot x, \tag{2.1.4}$$

so that equation (2.1.3b) can be rewritten as

$$v + \frac{\lambda}{\mu}(\nabla u + k_r(s)\nabla\psi) = 0. \tag{2.1.5}$$

By integrating the water volume conservation (2.1.3a) over K, we obtain

$$\int_K \partial_t(\phi s)\,\mathrm{d}x + \int_K \nabla\cdot v\,\mathrm{d}x = 0.$$

By applying the Stokes theorem, we get

$$\int_K \partial_t(\phi s)\,\mathrm{d}x + \int_{\partial K} v\cdot\nu_K\,\mathrm{d}\gamma = 0,$$

where ν_K stands for the outward normal vector to ∂K. Then, we write the previous equation at time t^n and discretizing the time partial derivative by an implicit Euler scheme, we obtain the semi-discretization

$$\int_K \phi\frac{s^n - s^{n-1}}{\Delta t^n}\,\mathrm{d}x + \int_{\partial K} v^n\cdot\nu_K\,\mathrm{d}\gamma = 0,$$

where $v^n = v(x,t^n)$. The second integral can be split into

$$\int_{\partial K} v^n\cdot\nu_K\,\mathrm{d}\gamma = \sum_{\sigma\in\mathscr{E}_K}\int_\sigma v^n\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma.$$

Let us look for an approximation $F^n_{K,\sigma}$ of the elementary flux

$$\int_\sigma v^n\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma = \int_\sigma \lambda\nabla u\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma + \frac{1}{\mu}\int_\sigma \lambda k_r(s^n)\nabla\psi\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma. \tag{2.1.6}$$

If $\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\rm ext}^{\rm N}$, by virtue of the boundary conditions (2.1.3e), a natural choice is

$$F^n_{K,\sigma} = \frac{1}{\Delta t^n}\int_{t^{n-1}}^{t^n}\mathrm{d}t\int_\sigma q^{\rm N}\,\mathrm{d}\gamma. \tag{2.1.7}$$

Let us now consider the cases $\sigma = K|L\in\mathscr{E}_K\cap\mathscr{E}_{\rm int}$ and $\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\rm ext}$. We define the transmissibilities $(A_\sigma)_{\sigma\in\mathscr{E}}$ by

$$A_\sigma = \begin{cases} \dfrac{m_\sigma}{\mu}\dfrac{\lambda_K\lambda_L}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} & \text{if}\quad \sigma = K|L\in\mathscr{E}_K\cap\mathscr{E}_{\rm int},\\ \dfrac{m_\sigma}{\mu}\dfrac{\lambda_K}{d_{K,\sigma}} & \text{if}\quad \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\rm ext}, \end{cases} \tag{2.1.8}$$

where m_σ is the $(d-1)$-Lebesgue measure of the edge $\sigma\in\mathscr{E}$, $\lambda_K = \frac{1}{m_K}\int_K\lambda\,\mathrm{d}x$ and

$$d_{K,\sigma} = \begin{cases} |x_K - x_L| & \text{if } \sigma = K\cap L\in\mathscr{E}_{\rm int},\\ \mathrm{dist}(x_K,\sigma) & \text{if } \sigma\in\mathscr{E}_{\rm ext}. \end{cases} \tag{2.1.9}$$

The TPFA scheme is then defined thanks to the approximations

$$-\frac{1}{\mu}\int_\sigma \lambda\nabla u\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma \simeq \begin{cases} A_\sigma(u_K - u_L) & \text{if } \sigma = K|L\in\mathscr{E}_K\cap\mathscr{E}_{\rm int},\\ A_\sigma(u_K - u_\sigma) & \text{if } \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\rm ext}^{\rm D}, \end{cases} \tag{2.1.10}$$

and

$$-\frac{1}{\mu}\int_\sigma \lambda k_r(s^n)\nabla\psi\cdot\nu_{K,\sigma}\,\mathrm{d}\gamma \simeq \begin{cases} A_\sigma k_{r,\sigma}^{n,\mathrm{up}}(\psi_K-\psi_L) & \text{if } \sigma=K|L\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{int}},\\ A_\sigma k_{r,\sigma}^{n,\mathrm{up}}(\psi_K-\psi_\sigma) & \text{if } \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}\end{cases} \tag{2.1.11}$$

where the edge relative permeability is upwinded according to the sign of the difference in ψ, viz.,

$$k_{r,\sigma}^{n,\mathrm{up}} = \begin{cases} k_r(s_K^n) & \text{if } \psi_K-\psi_{K,\sigma}\geqslant 0,\\ k_r(s_{K\sigma}^n) & \text{if } \psi_K-\psi_{K,\sigma}<0,\end{cases} \tag{2.1.12}$$

in which $s_{K\sigma}^n$ stands for the mirror value of s_K^n as defined in (2.1.2) inside the domain. At the boundary,

$$s_{K\sigma}^n = \begin{cases} s_L^n & \text{if } \sigma=K|L\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{int}},\\ s_\sigma^n=\mathscr{S}(u_\sigma^n) & \text{if } \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}},\end{cases}$$

Gathering the previous approximations, we obtain the approximate flux

$$F_{K,\sigma}^n = \begin{cases} A_\sigma[u_K^n-u_{K\sigma}^n+k_{r,\sigma}^{n,\mathrm{up}}(\psi_K-\psi_{K,\sigma})] & \text{if } \sigma\in\mathscr{E}_K\cap(\mathscr{E}_{\mathrm{int}}\cup\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}),\\ \dfrac{1}{\Delta t^n}\displaystyle\int_{t^{n-1}}^{t^n}\mathrm{d}t\int_\sigma q^{\mathrm{N}}\,\mathrm{d}\gamma & \text{if } \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}.\end{cases} \tag{2.1.13}$$

This numerical flux is then consistent under the orthogonality assumption on the mesh in the sense that, for any regular function $\varphi\in\mathcal{C}_c^\infty(\Omega)$, we have

$$\left|-\frac{1}{\mu}\int_\sigma \lambda\nabla\varphi\cdot\nu_{K,\sigma}d\sigma - F_{K,\sigma}\right| \leqslant m_\sigma\mathcal{C}_{\varphi,\lambda,\mu}h$$

for all $K\in\mathscr{T}$, $\mathcal{C}_{\varphi,\lambda,\mu}\in\mathbb{R}_+$ depends on φ, λ and μ and $h=\max\{\mathrm{diam}(K),K\in\mathscr{T}\}$. Note that, from equation (2.1.13), the local conservation property

$$F_{K,\sigma}+F_{L,\sigma}=0,$$

holds for $\sigma=K|L$. We also point out that, for an inner or Dirichlet face, the flux (2.1.13) can be recast under the form

$$F_{K,\sigma}^n = A_\sigma[u_K^n-u_{K\sigma}^n+k_r(s_K^n)(\psi_K-\psi_{K\sigma})^+ + k_r(s_{K,\sigma}^n)(\psi_K-\psi_{K\sigma})^-] \tag{2.1.14}$$

with $a^+=\max(a,0)$ and $a^-=\min(a,0)$. Our discrete equation finally reads

$$m_K\phi_K\frac{s_K^n-s_K^{n-1}}{\Delta t^n}+\sum_{\sigma\in\mathscr{E}_K}F_{K,\sigma}^n=0, \tag{2.1.15a}$$

$$s_K^n-\mathscr{S}(u_K^n)=0, \tag{2.1.15b}$$

for all $K\in\mathscr{T}$ and $n\geqslant 1$, where $F_{K,\sigma}^n$ is defined by (2.1.13). This discretized scheme has been studied in [75] where a convergence proof is stated.

Pressure-saturation formulation. Following the same procedure, the discretized scheme for the pressure-saturation formulation system (1.2.13) is

$$m_K \phi_K \frac{s_K^n - s_K^{n-1}}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}^n = 0, \tag{2.1.16a}$$

$$s_K^n - \mathcal{S}(p_K^n) = 0, \tag{2.1.16b}$$

with

$$F_{K,\sigma}^n = \begin{cases} A_\sigma k_{r,\sigma}^{n,\mathrm{up}} [\vartheta_K^n - \vartheta_{K\sigma}^n] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\mathrm{int}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} \int_\sigma q^{\mathrm{N}} \, \mathrm{d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}, \end{cases} \tag{2.1.17}$$

where the edge relative permeability

$$k_{r,\sigma}^{n,\mathrm{up}} = \begin{cases} k_r(s_K^n) & \text{if } \vartheta_K^n - \vartheta_{K\sigma}^n \geqslant 0, \\ k_r(s_{K\sigma}^n) & \text{if } \vartheta_K^n - \vartheta_{K\sigma}^n < 0, \end{cases} \tag{2.1.18}$$

is now upwinded according to the sign of the difference in the hydraulic head

$$\vartheta^n = p^n + \psi. \tag{2.1.19}$$

In Chapter 3 we study this discretized scheme establishing the existence of a unique solution for this scheme and providing a rigorous mathematical convergence proof.

2.2 Parametrization of the characteristic laws

Let us recall the motivations behind the choice of the primary variable τ. Choosing the pressure as the primary variable is known to be inefficient for dry soils $s \ll 1$. On the other hand, the knowledge of the saturation is not sufficient to describe the pressure in saturated regions where $s = 1 - s_{\mathrm{rn}}$. This motivated the introduction of schemes based on variable switching between s and p in [58,81]. Indeed, throughout this thesis, we adopt a formulation that is close to the original variable switch since our new variable τ behaves either as s or as p (or u) up to a linear function.

Pressure-saturation formulation. Our approach is based on [29] and can be seen as a reformulation of the variable switch which makes its implementation much easier. The idea is to choose a parametrization of the graph $\{p, \mathcal{S}(p)\}$, i.e. to choose two functions

$$\mathfrak{s} : I \to [s_{\mathrm{rw}}, 1 - s_{\mathrm{rn}}], \qquad \mathfrak{p} : I \to \mathbb{R},$$

such that

$$\mathfrak{s}(\tau) = \mathcal{S}(\mathfrak{p}(\tau)), \qquad 0 < \mathfrak{s}'(\tau) + \mathfrak{p}'(\tau) < \infty,$$

for all $\tau \in I \subset \mathbb{R}$. Such a parametrization is not unique: one can for instance choose $I = \mathbb{R}$, $\mathfrak{p} = \mathrm{Id}$ and $\mathfrak{s} = \mathcal{S}$, or $\mathfrak{p} = (\mathrm{Id} + \mathcal{S})^{-1}$ and $\mathfrak{s} = (\mathrm{Id} + \mathcal{S}^{-1})^{-1}$. The difficulty lies in finding the optimal formulation. In the pressure-saturation formulation, we take $I = (s_{\mathrm{rw}}, +\infty)$ and

$$(\mathfrak{s}(\tau),\, \mathfrak{p}(\tau)) = \begin{cases} (\tau,\, \mathcal{S}^{-1}(\tau)) & \text{if } \tau \leqslant s_{\mathrm{s}}, \\ \Big(\mathcal{S}\Big(p_{\mathrm{s}} + \dfrac{\tau - s_{\mathrm{s}}}{\mathcal{S}'(p_{\mathrm{s}}^-)}\Big),\, p_{\mathrm{s}} + \dfrac{\tau - s_{\mathrm{s}}}{\mathcal{S}'(p_{\mathrm{s}}^-)}\Big) & \text{if } \tau \geqslant s_{\mathrm{s}}, \end{cases} \tag{2.2.1}$$

where $\mathcal{S}'(p_{\rm s}^-)$ denotes the limit of $\mathcal{S}'(p)$ as p tends to $p_{\rm s}$ from below. $(p_{\rm s}, \tau_{\rm s})$ is the inflection point of $\mathcal{S}$, which is convex and then concave in both the Brooks-Corey model and van Genuchten-Mualem settings, cf. Section 1.2.3. Thus both $\mathfrak{s}$ and $\mathfrak{p}$ are C^1 and concave, and even C^2 if $\mathcal{S}$ is given by (1.2.27). Moreover, for all $p \in \mathbb{R}$, there exists a unique $\tau \in (s_{\rm rw}, +\infty)$ such that $(p, \mathcal{S}(p)) = (\mathfrak{p}(\tau), \mathfrak{s}(\tau))$.

The discretized equations (2.1.16)–(2.1.18) become

$$m_K \phi_K \frac{\mathfrak{s}(\tau_K^n) - \mathfrak{s}(\tau_K^{n-1})}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}^n = 0, \tag{2.2.2}$$

with

$$F_{K,\sigma}^n = \begin{cases} A_\sigma k_{r,\sigma}^{n,\rm up} \left[\vartheta(\tau_K^n) - \vartheta(\tau_{K\sigma}^n)\right] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm D}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} {\rm d}t \int_\sigma q^{\rm N} \, {\rm d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N} \end{cases} \tag{2.2.3}$$

where

$$k_{r,\sigma}^{n,\rm up} = \begin{cases} k_r(\mathfrak{s}(\tau_K^n)) & \text{if } \vartheta(\tau_K^n) - \vartheta(\tau_{K\sigma}^n) \geqslant 0, \\ k_r(\mathfrak{s}(\tau_{K,\sigma}^n)) & \text{if } \vartheta(\tau_K^n) - \vartheta(\tau_{K\sigma}^n) < 0, \end{cases} \tag{2.2.4}$$

is the upwinded edge relative permeability and

$$\vartheta(\tau_K^n) = \mathfrak{p}(\tau_K^n) + \psi_K, \qquad \vartheta(\tau_{K\sigma}^n) = \mathfrak{p}(\tau_{K\sigma}^n) + \psi_{K\sigma}, \tag{2.2.5}$$

are the hydraulic heads.

Kirchhoff transform-saturation formulation. Similarly to the above approach, the idea is to consider a parametrization of the graph $\{u, \mathscr{S}(u)\}$ by means of two functions

$$\mathfrak{s} : I \to [s_{\rm rw}, 1 - s_{\rm rn}] \quad \text{and} \quad \mathfrak{u} : I \to \mathbb{R}$$

such that $\mathfrak{s}(\tau) = \mathscr{S}(\mathfrak{u}(\tau))$ for all $\tau \in I \subset \mathbb{R}$. Again, even subject to the speed-control condition

$$\mathfrak{s}'(\tau) + \mathfrak{u}'(\tau) = 1,$$

such a parametrization is not unique. In the spirit of the variable-switching method, we set $I = (s_{\rm rw}, +\infty)$ and

$$(\mathfrak{s}(\tau),\, \mathfrak{u}(\tau)) = \begin{cases} (\tau,\, \mathscr{S}^{-1}(\tau)) & \text{if } \tau \leqslant \tau_{\rm s}, \\ \Big(\mathscr{S}\Big(u_{\rm s} + \dfrac{\tau - s_{\rm s}}{\mathscr{S}'(u_{\rm s})}\Big),\, u_{\rm s} + \dfrac{\tau - s_{\rm s}}{\mathscr{S}'(u_{\rm s})}\Big) & \text{if } \tau \geqslant \tau_{\rm s}, \end{cases} \tag{2.2.6}$$

with $\tau_{\rm s} = s_{\rm s}$ the switch value obtained by imposing $s_s = \mathscr{S}(u_{\rm s})$. We recall that this Kirchhoff transform-saturation formulation, and consequently this parametrization, is not always usable since, depending on the chosen model for the characteristic laws, the Kirchhoff transform may not be analytically computed. In our work this parametrization using the Kirchhoff transform formulation is used only when the Brooks-Corey model is employed.

The discretized equations (2.1.17)–(2.1.18) become

$$m_K \phi_K \frac{\mathfrak{s}(\tau_K^n) - \mathfrak{s}(\tau_K^{n-1})}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}^n = 0, \tag{2.2.7}$$

with

$$F_{K,\sigma}^n = \begin{cases} A_\sigma[\mathfrak{u}(\tau_K^n) - \mathfrak{u}(\tau_{K\sigma}^n) + k_{r,\sigma}^{n,\mathrm{up}}(\psi_K - \psi_{K\sigma})] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\mathrm{int}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}), \\ \dfrac{1}{\Delta t^n}\displaystyle\int_{t^{n-1}}^{t^n} \mathrm{d}t \int_\sigma q^{\mathrm{N}}\,\mathrm{d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathcal{E}_{\mathrm{ext}}^{\mathrm{N}}, \end{cases} \tag{2.2.8}$$

where

$$k_{r,\sigma}^{n,\mathrm{up}} = \begin{cases} k_r(\mathfrak{s}(\tau_K^n)) & \text{if } \psi_K - \psi_{K\sigma} \geqslant 0, \\ k_r(\mathfrak{s}(\tau_{K\sigma}^n)) & \text{if } \psi_K - \psi_{K\sigma} < 0. \end{cases} \tag{2.2.9}$$

Remark 2.2.1. *If we define $s_K^n = \mathfrak{s}(\tau_K^n)$ and $p_K^n = \mathfrak{p}(\tau_K^n)$ —resp. $u_K^n = \mathfrak{u}(\tau_K^n)$— then the equation $s_K^n = \mathcal{S}(p_K^n)$ —resp. $s_K^n = \mathscr{S}(u_K^n)$— is automatically satisfied.*

2.3 Iterative solver for the nonlinear system

2.3.1 Classical Newton-Raphson method

At each time-step, both systems (2.2.7)–(2.2.8) and (2.2.2)–(2.2.3) can be rewritten abstractly as

$$\mathcal{F}_n(\boldsymbol{\tau}^n) := (f_K^n(\boldsymbol{\tau}^n))_{K\in\mathscr{T}} = \mathbf{0}, \tag{2.3.1}$$

which consists of $\#\mathscr{T}$ nonlinear equations, each component of which is defined as

$$f_K^n(\boldsymbol{\tau}^n) = m_K\phi_K\frac{\mathfrak{s}(\tau_K) - \mathfrak{s}(\tau_K^{n-1})}{\Delta t^n} + \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n, \qquad K \in \mathscr{T}. \tag{2.3.2}$$

Solving (2.3.1) with Newton's method amounts to constructing a sequence $\{\boldsymbol{\tau}^{n,k}\}_{k\geqslant 0}$ defined by

$$\boldsymbol{\tau}^{n,0} = \boldsymbol{\tau}^{n-1}, \tag{2.3.3a}$$

$$\boldsymbol{\tau}^{n,k+1} = \boldsymbol{\tau}^{n,k} - [\mathbf{J}_{\mathcal{F}_n}(\boldsymbol{\tau}^{n,k})]^{-1}\mathcal{F}_n(\boldsymbol{\tau}^{n,k}), \tag{2.3.3b}$$

where

$$\mathbf{J}_{\mathcal{F}_n}(\boldsymbol{\tau}^{n,k}) = \frac{\partial\mathcal{F}_n}{\partial\boldsymbol{\tau}}(\boldsymbol{\tau}^{n,k}) = \left(\frac{\partial f_K^n}{\partial\tau_L^n}(\boldsymbol{\tau}^{n,k})\right)_{K\in\mathscr{T},L\in\mathscr{T}}.$$

is the Jacobian matrix of $\mathcal{F}_n$ at $\boldsymbol{\tau}^{n,k}$, assumed to be nonsingular. Let us now detail the procedure for both formulations.

Kirchhoff transform-saturation formulation. The left-hand side of the discrete equation (2.2.8) can be transformed, for all $K \in \mathscr{T}$, into

$$\begin{aligned} f_K^n(\boldsymbol{\tau}) = {} & m_K\phi_K\frac{\mathfrak{s}(\tau_K) - \mathfrak{s}(\tau_K^{n-1})}{\Delta t^n} \\ & + \sum_{\sigma\in\mathscr{E}_{\mathrm{int}}\cup\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}} A_\sigma[\mathfrak{u}(\tau_K) - \mathfrak{u}(\tau_{K\sigma}) + k_r(\mathfrak{s}(\tau_K))(D_{K,\sigma}\psi)^+ + k_r(\mathfrak{s}(\tau_{K\sigma}))(D_{K,\sigma}\psi)^-] \\ & + \sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}} \frac{1}{\Delta t^n}\int_{t^{n-1}}^{t^n}\mathrm{d}t\int_\sigma q^{\mathrm{N}}\,\mathrm{d}\gamma, \end{aligned} \tag{2.3.4}$$

where

$$D_{K,\sigma}\psi = \psi_K - \psi_{K\sigma}. \tag{2.3.5}$$

For this scheme, the Jacobian matrix is defined by

$$[\mathbf{J}_{\mathcal{F}_n}]_{K,K}(\boldsymbol{\tau}) = \frac{m_K\phi_K}{\Delta t^n}\mathfrak{s}'(\tau_K) + \sum_{\sigma\in\mathscr{E}_{\rm int}\cup\mathscr{E}_{\rm ext}^{\rm D}} A_\sigma[\mathfrak{u}'(\tau_K) + k_r'(\mathfrak{s}(\tau_K))\mathfrak{s}'(\tau_K)(D_{K,\sigma}\psi)^+], \tag{2.3.6a}$$

$$[\mathbf{J}_{\mathcal{F}_n}]_{K,L}(\boldsymbol{\tau}) = A_\sigma[-\mathfrak{u}'(\tau_L) + k_r'(\mathfrak{s}(\tau_L))\mathfrak{s}'(\tau_L)(D_{K,\sigma}\psi)^-], \tag{2.3.6b}$$

for all $L\in\mathscr{T}$ such that $\sigma = K|L\in\mathscr{E}_K$.

Pressure-saturation formulation. The left-hand side of the discrete equation (2.2.3) can be transformed, for all $K\in\mathscr{T}$, into

$$\begin{aligned} f_K^n(\boldsymbol{\tau}^n) = {}& m_K\phi_K\frac{\mathfrak{s}(\tau_K^n)-\mathfrak{s}(\tau_K^{n-1})}{\Delta t^n} \\ &+ \sum_{\sigma\in\mathscr{E}_{\rm int}\cup\mathscr{E}_{\rm ext}^{\rm D}} A_\sigma[k_r(\mathfrak{s}(\tau_K^n))(D_{K,\sigma}\vartheta^n)^+ + k_r(\mathfrak{s}(\tau_{K\sigma}^n))(D_{K,\sigma}\vartheta^n)^-] \\ &+ \sum_{\sigma\in\mathscr{E}_{\rm ext}^{\rm N}} \frac{1}{\Delta t^n}\int_{t^{n-1}}^{t^n}\mathrm{d}t\int_\sigma q^{\rm N}\,\mathrm{d}\gamma, \end{aligned} \tag{2.3.7}$$

where

$$D_{K,\sigma}\vartheta^n = \vartheta_K^n - \vartheta_{K\sigma}^n. \tag{2.3.8}$$

For this scheme, the Jacobian matrix is defined by

$$\begin{aligned} [\mathbf{J}_{\mathcal{F}_n}]_{K,K}(\boldsymbol{\tau}) = {}& \frac{m_K\phi_K}{\Delta t^n}\mathfrak{s}'(\tau_K) \\ &+ \sum_{\sigma\in\mathscr{E}_{\rm int}\cup\mathscr{E}_{\rm ext}^{\rm D}} A_\sigma[k_{r,\sigma}^{\rm up}\mathfrak{p}'(\tau_K) + k_r'(\mathfrak{s}(\tau_K))\mathfrak{s}'(\tau_K)(D_{K,\sigma}\vartheta)^+], \end{aligned} \tag{2.3.9a}$$

$$[\mathbf{J}_{\mathcal{F}_n}]_{K,L}(\boldsymbol{\tau}) = A_\sigma[-k_{r,\sigma}^{\rm up}\mathfrak{p}'(\tau_{K\sigma}) + k_r'(\mathfrak{s}(\tau_{K\sigma}))\mathfrak{s}(\tau_{K\sigma})(D_{K,\sigma}\vartheta)^-], \tag{2.3.9b}$$

for all $L\in\mathscr{T}$ such that $\sigma = K|L\in\mathscr{E}_K$.

2.3.2 Enhancements of the Newton-Raphson method

The Newton's algorithm we have implemented is detailed in Algorithm 1. It includes some additional functionalities, listed below, whose aim is to treat the difficulties presented by the constitutive laws.

- `truncate`()
  Since $\mathcal{F}_n$ is not necessarily C^1 ($\mathcal{S}_{\rm BC}$ is not C^1 in the Brooks-Corey case), following [93,135], the Newton increment is truncated near the inflection point $s_{\rm s}$.

- `decreaseDeltaTime`() and `increaseDeltaTime`()
  In our numerical tests, we increase the time step in such a way that

$$\Delta t^{n+1} = \min(\Delta t_{\max},\, \alpha_{\Delta t}^+\Delta t^n), \qquad \alpha_{\Delta t}^+ > 1,$$

Initialization:
$k = 0$;
$\boldsymbol{\tau}^{n,0} = \boldsymbol{\tau}^{n-1}$;
while $[\|\mathcal{F}_n(\boldsymbol{\tau}^{n,k})\|_\infty \geqslant \epsilon$ and $k \leqslant k_{\max}]$ **do**
  solve $\mathbf{J}(\boldsymbol{\tau}^{n,k})\boldsymbol{\delta}^{n,k} + \mathcal{F}_n(\boldsymbol{\tau}^{n,k}) = \mathbf{0}$;
  for $K \in \mathscr{T}$ **do**
    `truncate`();
    $\tau_K^{n,k+1} = \max(s_{\rm rw}, \tau_K^{n,k} + \delta_K^{n,k})$;
  end
  $k = k + 1$;
end
if $k > k_{\max}$ **then**
  `decreaseDeltaTime`();
else
  $\boldsymbol{\tau}^n = \boldsymbol{\tau}^{n,k}$;
  $n = n + 1$;
  `increaseDeltaTime`();
end

Algorithm 1: Practical resolution of the system $\mathcal{F}_n(\boldsymbol{\tau}^n) = \mathbf{0}$.

and decrease it in such a way that

$$\Delta t^{n+1} = \max(\alpha_{\Delta t}^- \Delta t^n,\, \Delta t_{\min}), \qquad \alpha_{\Delta t}^- < 1.$$

If $\Delta t_{\min}$ is reached, the simulation stops.

Moreover, an approximation of the law of the relative permeability k_r in the van Genuchten-Mualem case (1.2.27) is required because it presents very large derivative's values, which can be equal to ∞, for $s \to 1 - s_{\rm rn}$. We detail the approximation in the following. Let us now explain why we introduced the truncation procedure. When we have presented the constitutive laws for the Brooks-Corey and van Genuchten-Mualem models, we have pointed out that $\mathcal{S}_{\rm BC}$, $\mathscr{S}_{\rm BC}$ are not C^1 and that the derivative of the relative permeability function $k_{r\rm vGM}$ blows up at $s = 1 - s_{\rm rn}$. Thus, as well as the possibly very large value of the derivative of k_r affects the conditioning of the Jacobian matrix, during Newton's iterations, the presence of corner points may lead to wrong gradient directions, resulting in a non-convergence of the method. Here we show how we have handled these difficulties.

Truncation method to treat corner points. Let us assume that the function $f(\tau)$ has a corner point in $\tau = \tau_s$. The truncation procedure is activated in the two following cases.

1. $\tau_K^{n,k} > \tau_s$ and the expected $\tau_K^{n,k+1} \leqslant \tau_s$. Then, we set $\tau_K^{n,k+1} = \tau_s - \epsilon_\delta$, where $0 < \epsilon_\delta \ll 1$ is a fixed threshold. This is illustrated in Figure 2.2.

2. $\tau_K^{n,k} < \tau_s$ and the expected $\tau_K^{n,k+1} > \tau_s$. Then, we set $\tau_K^{n,k+1} = \tau_s + \epsilon_\delta$. This case is the symmetric of the previous one, as illustrated in Figure 2.3.

The procedure is summarized in the pseudocode Algorithm 2.

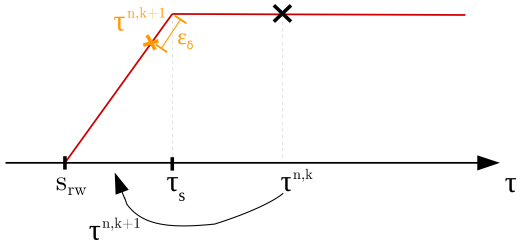


Figure 2.2: Case $\tau_K^{n,k} > \tau_s$, $\tau_K^{n,k+1} \leqslant \tau_s$.

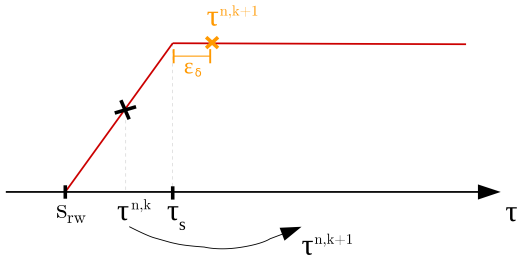


Figure 2.3: Case $\tau_K^{n,k} < \tau_s$, $\tau_K^{n,k+1} > \tau_s$.

$\boldsymbol{\tau}^{n,k+1} = \boldsymbol{\tau}^{n,k} - \mathbf{J}_{\mathcal{F}_n}^{-1}(\boldsymbol{\tau}^{n,k})\mathcal{F}_n(\boldsymbol{\tau}^{n,k})$;
for $K \in \mathscr{T}$ **do**
  if $(\tau_K^{n,k} - \tau_s)(\tau_K^{n,k+1} - \tau_s) \leqslant 0$ **then**
    if $\tau_K^{n,k} > \tau_s$ **then**
      $\tau_K^{n,k+1} = \tau_s - \epsilon_\delta$;
    else
      $\tau_K^{n,k+1} = \tau_s + \epsilon_\delta$;
    end
end

Algorithm 2: Pseudocode for the truncation procedure.

Chord slope method using a second degree polynomial. The idea is to smooth the relative permeability law in the area where the derivative blows up by approximating it locally with a polynomial. Its form is chosen in order to allow us to preserve the convexity of the real law. Let us consider $k_r(s)$ whose derivative tends to infinite when $s \to 1 - s_{\mathrm{rn}}$. We fix a value $s_{\mathrm{lim}} < 1$ and, to interpolate the law between this point and $s = 1$, we use a quadratic polynomial, namely,

$$\wp(s) = \frac{k_r''(s_{\mathrm{lim}})}{2}(s - s_{\mathrm{lim}})^2 + k_r'(s_{\mathrm{lim}})(s - s_{\mathrm{lim}}) + k_r(s_{\mathrm{lim}}).$$

Consequently the approximated law reads

$$\widetilde{k}_r(s) = \begin{cases} k_r(s) & \text{if } s \leqslant s_{\mathrm{lim}} \\ \wp(s) & \text{if } s > s_{\mathrm{lim}} \end{cases}.$$

In Figure 2.4 we can see the profile of the relative permeability law and of the modified one. This technique has already been applied, for instance, in [81]. But, contrary to us, in this work, the approximation is applied not only to the relative permeability but also to the capillary pressure law.

In [18] we propose the same approximation with the only difference that interpolations are updated during the iterations and the converge is reached with the original laws.

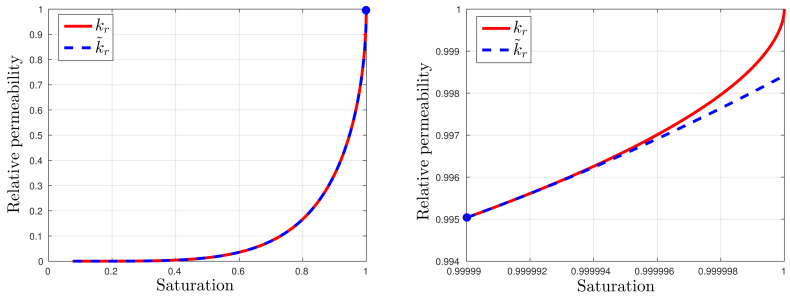


Figure 2.4: Shape of the relative permeability law and of the modified one for $s_{\text{lim}} = 1 - s_{\text{rw}} - 10^{-5}$.

2.4 Numerical results

We now present different tests for the numerical resolution of the Richards equation in the homogeneous case using the pressure-saturation formulation and the Kirchhoff transform-saturation one. We first test the robustness of the Kirchhoff transform-saturation formulation under different test settings. While the linearization property of the Kirchhoff-transform-saturation formulation is advantageous, it is not always the best option. Firstly, the Kirchhoff transform cannot always be calculated analytically, e.g., in the case of the van Genuchten-Mualem model. Secondly, when considering a domain composed of different rock types (heterogeneous case), transmissivity conditions are imposed at the interfaces between different lithologies to ensure pressure continuity. In this case, it is more natural to work directly with pressures. This motivates our interest in working in pressure-saturation formulation. With this in mind, a test comparing the use of the two formulations is presented, followed by other tests aimed at showing the advantage of using the parametrization technique rather than solving the problem solely in the pressure or saturation variable.

2.4.1 Kirchhoff transform-saturation formulation

For the numerical validation of the Kirchhoff transform-saturation formulation (2.2.7)–(2.2.8) , using the parametrized laws (2.2.6), we consider two test cases inspired from those proposed in [29]. For the simulations we take the following parameters: $\epsilon = 10^{-12}$, $i_{\max} = 500$, $\epsilon_\delta = 10^{-7}$. The time-step is taken constant during each simulation.

2.4.1.1 Robustness

Let us consider a two-dimensional domain $\Omega = [0,1] \times [0,1]$ (in meters) which is initially very dry with $s_0 = 10^{-6}$. Water is injected trough the portion of the upper boundary $\Gamma^{\text{D}} = \{(x,y) \,|\, x \in [0, 0.3], y = 1\}$ (in meters) imposing $u_{\text{D}} = 1$. A zero flux boundary condition is prescribed on Γ^{N}. The configuration of the domain is shown in Figure 2.5. The goal of this test is challenge the robustness of the scheme. For this reason, some parameters are taken equal to one for seek of

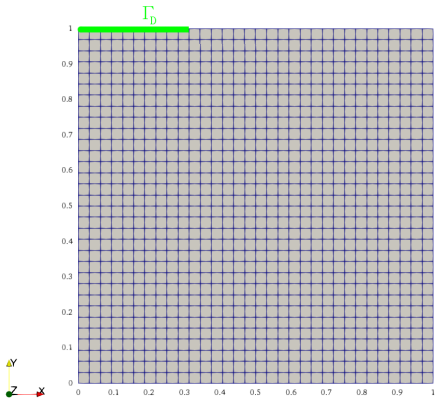


Figure 2.5: Configuration of the domain Ω for the filling test.

simplicity: the gravity vector is chosen in such a way that $g = -e_y$, the porosity of the medium and its permeability are considered equal to one ($\phi = 1,\ \lambda = 1\mathrm{m}^2$) as well as the wetting phase viscosity ($\mu = 1\mathrm{Pa}\cdot\mathrm{s}$). We consider the Brooks-Corey model defined in (1.2.30) with $s_{\mathrm{rn}} = s_{\mathrm{rw}} = 0$. In order to test the robustness of the parametrization, we set $p_b = 10^{-2}$ and let the parameter n take values in $\{1, 2, 4, 8, 16\}$. The curves of the characteristic laws, evaluated for each n, are reported in Figure 2.6. For each value of n, we compute a reference solution denoted by $(\boldsymbol{\tau}_{\mathrm{ref}})$ taking for the tolerance

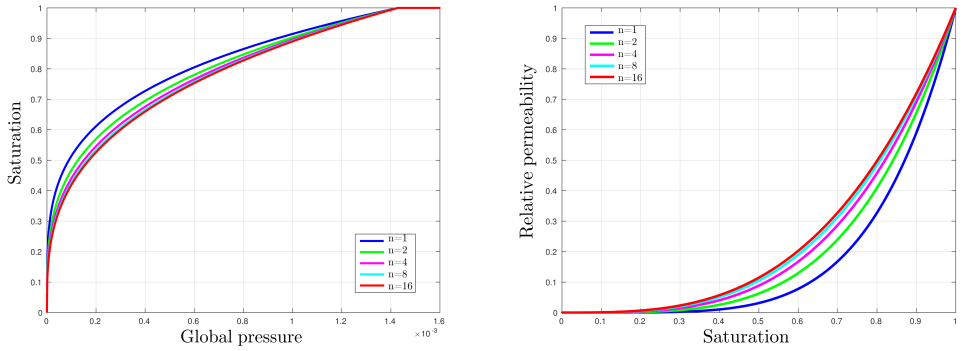


Figure 2.6: Profiles of the saturation-Kirchhoff transform relationship and relative permeability law for $n \in \{1, 2, 4, 8, 16\}$.

of Newton's method $\epsilon_{n,\mathrm{ref}} = 10^{-12}$. Then, for each value of n, we perform calculations taking as tolerance $\epsilon \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}\}$. We call these solutions $(\boldsymbol{\tau}_{n,\epsilon})$ and measure the average number of iterations per time-step for each value of ϵ. Then we compute the deviation of the observable variables u and s from the reference solution. The error committed is measured by the quantities:

$$\mathtt{err}^{u}_{n,\epsilon} = \frac{\|\mathfrak{u}(\boldsymbol{\tau}_{n,\epsilon}) - \mathfrak{u}(\boldsymbol{\tau}_{n,\mathrm{ref}})\|_{L^\infty(0,T;L^\infty(\Omega))}}{\|\mathfrak{u}(\boldsymbol{\tau}_{n,\mathrm{ref}})\|_{L^\infty(0,T;L^\infty(\Omega))}}, \tag{2.4.1a}$$

$$\mathtt{err}^{s}_{n,\epsilon} = \frac{\|\mathfrak{s}(\boldsymbol{\tau}_{n,\epsilon}) - \mathfrak{s}(\boldsymbol{\tau}_{n,\mathrm{ref}})\|_{L^\infty(0,T;L^\infty(\Omega))}}{\|\mathfrak{s}(\boldsymbol{\tau}_{n,\mathrm{ref}})\|_{L^\infty(0,T;L^\infty(\Omega))}}. \tag{2.4.1b}$$

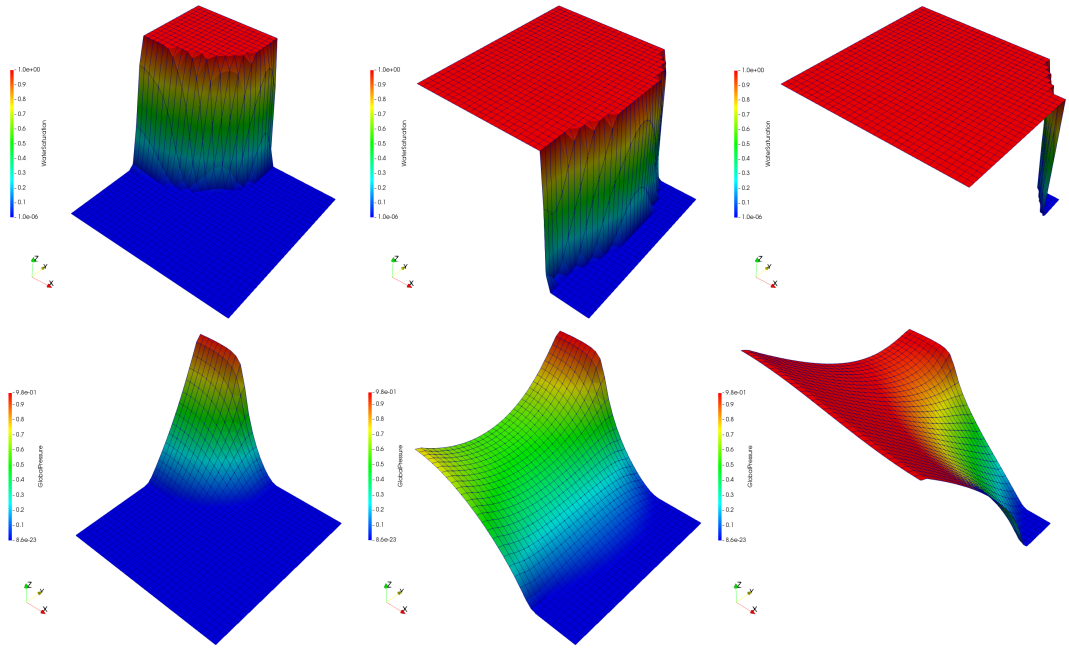


Figure 2.7: Profile of the saturation (top) and Kirchhoff transform (below) for $t \in \{0.1, 0.5, 0.7\}$ (in seconds).

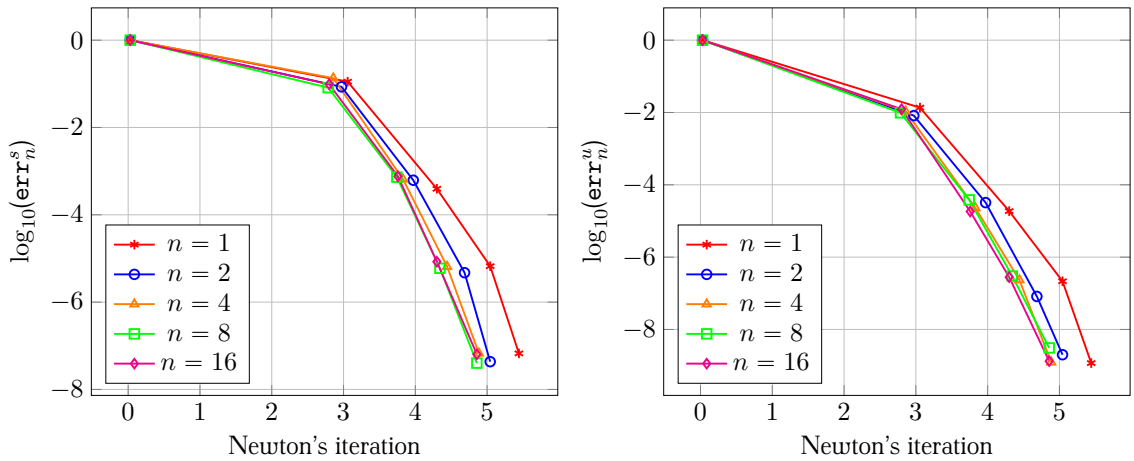


Figure 2.8: Relative errors given by (2.4.1) as functions of the average number of Newton iterations per time-step for each value of $\epsilon \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}\}$.

The simulation is performed on a mesh composed of 32×32 cells during a time interval $[0, T]$ with $T = 0.7\,\mathrm{s}$. The time-step is equal to $\Delta t = 0.01\,\mathrm{s}$. The truncation procedure is activated during Newton's iterations. The profile of saturation and Kirchhoff transform of the reference solution, using $n = 16$, at different moments of the simulation are reported in Figure 2.7. Finally the error profiles for $\mathtt{err}^u_{n,\epsilon}$ and $\mathtt{err}^s_{n,\epsilon}$ are reported in Figure 2.8. We can notice that the number of iterations required to reach the tolerance ϵ remains constant with respect to the parameter n.

Remark 2.4.1. *It is very important to underline that if we solve the problem without the variable switch, Newton's method does not converge because of the infinite derivative of the saturation-Kirchhoff transform curve at point* $(0, 0)$ *(see left panel in Figure 2.6).*

2.4.1.2 Filling the domain

In this test, we simulate the filling of an initially very dry domain of sand $\Omega = [0, 1] \times [0, 1]$ (in meters), see Figure 2.5. The hydraulic properties of this rock type are given in Table 2.1. The initial saturation is set to

$$s_0 = 10^{-6} \quad \text{in } \Omega. \tag{2.4.2}$$

The water is injected at pressure $u_{\mathrm{D}} = 500 > u_b = 454$ through a portion of the upper boundary $\Gamma^{\mathrm{D}} = \{(x, y) \,|\, x \in [0, 0.3], y = 1\}$ (in meters) and a no flux boundary condition is imposed on $\partial\Omega \backslash \Gamma^{\mathrm{D}}$. For simplicity the gravity vector is chosen in such a way that$\mu = 10^{-3}\,\mathrm{Pa}\cdot\mathrm{s}$, $\rho = 1\,\mathrm{Kg}\cdot\mathrm{m}^{-3}$ and water viscosity is set equal to $\mu = 10^{-3}\,\mathrm{Pa}\cdot\mathrm{s}$. The configuration of the domain is shown in Figure 2.5. The simulation is performed on a mesh composed of 32×32 cells during a time interval $[0, T]$ with $T = 1.76 \cdot 10^5\,\mathrm{s}$. The time-step is equal to $\Delta t = 4 \cdot 10^3\,\mathrm{s}$. The truncation procedure is activated during Newton's iterations.

	$1 - s_{\mathrm{rn}}$	s_{rw}	n	$\lambda\,[\mathrm{m}^2]$	$p_b\,[\mathrm{Pa}]$	ϕ
Sand	1.0	0	2.239	$6.3812 \cdot 10^{-12}$	$3.5036 \cdot 10^3$	0.3658

Table 2.1: Hydraulic properties of the porous medium.

The values of saturation and Kirchhoff transform obtained at different times are shown in Figure 2.9. During the simulation we have registered 4 iterations for Newton's method on average with a maximum of 17 iterations.

Figure 2.10 shows the evolution of the average Newton's convergence rate given, for a time-step n, by

$$\mathtt{CV}^n_{rate} = \frac{1}{N^n_{iter}} \sum_{k=0}^{N^n_{iter}-1} \frac{\log_{10} \left\| \mathcal{F}_n(\boldsymbol{\tau}^{n,k+1}) \right\|_\infty}{\log_{10} \left\| \mathcal{F}_n(\boldsymbol{\tau}^{n,k}) \right\|_\infty}.$$

Remark 2.4.2. *Notice that, just on the Dirichlet boundary at the beginning and then progressively in the whole domain, pressures are higher than the entry pressure* u_b *and the saturation-pressure relationship is a graph. If we solve the problem in Kirchhoff transform variable, Newton's method does not converge. On the other hand, is not possible to solve the problem in saturation variable because we have* $u > u_b$*. The problem can still be solved thanks to the use of the parametrization technique that permits to select the suitable primary variable for each cell of the mesh depending on its saturation-pressure value.*

In order to further illustrate the robustness of our approach, let us now perform the same simulation in just one time-step with $\Delta t = T = 1.76 \cdot 10^5\,\mathrm{s}$. To achieve the convergence, 60

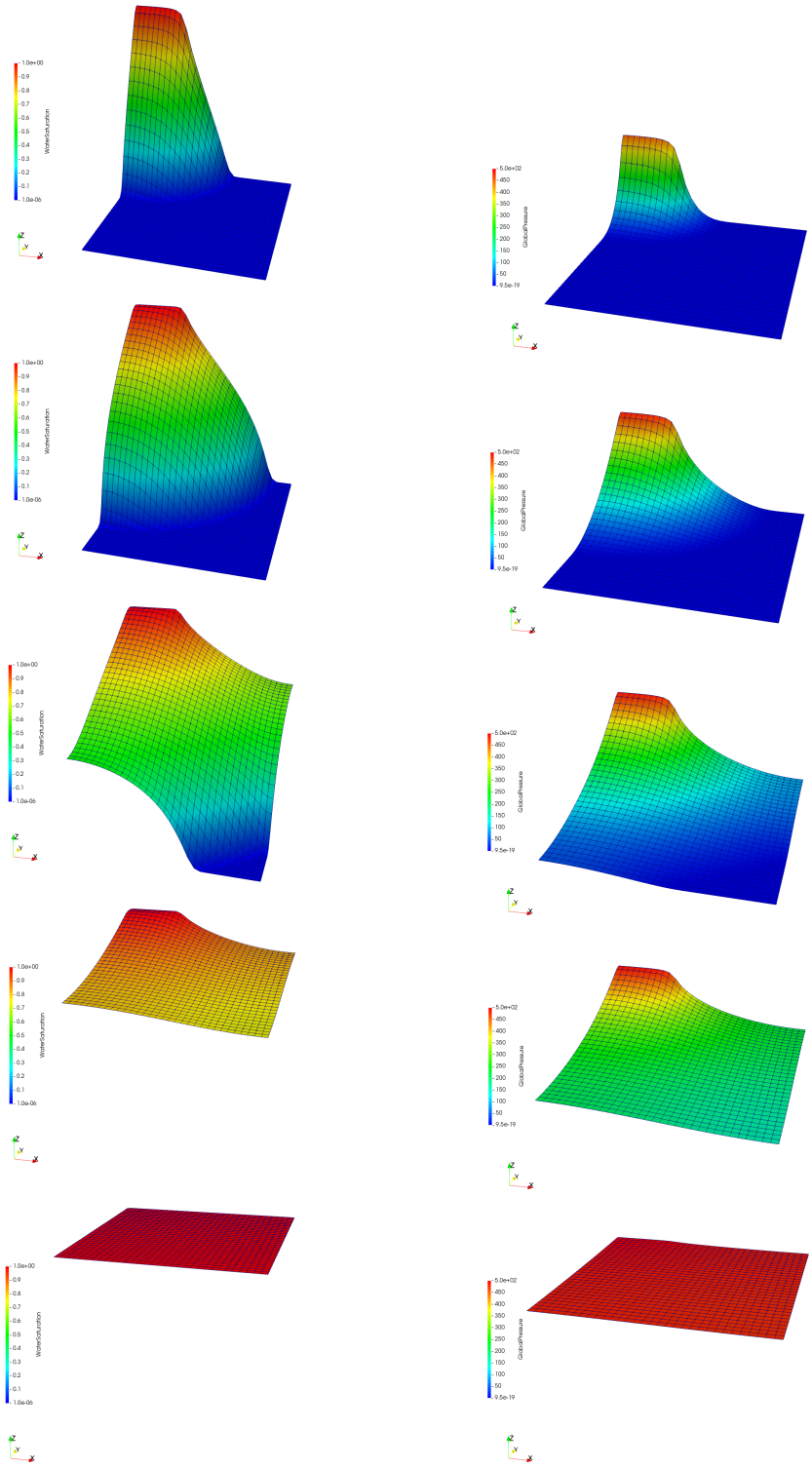


Figure 2.9: Saturation and Kirchhoff transform profiles, on the left and right columns respectively, for $t \in \{1 \cdot \Delta t, 5 \cdot \Delta t, 15 \cdot \Delta t, 25 \cdot \Delta t, T\}$

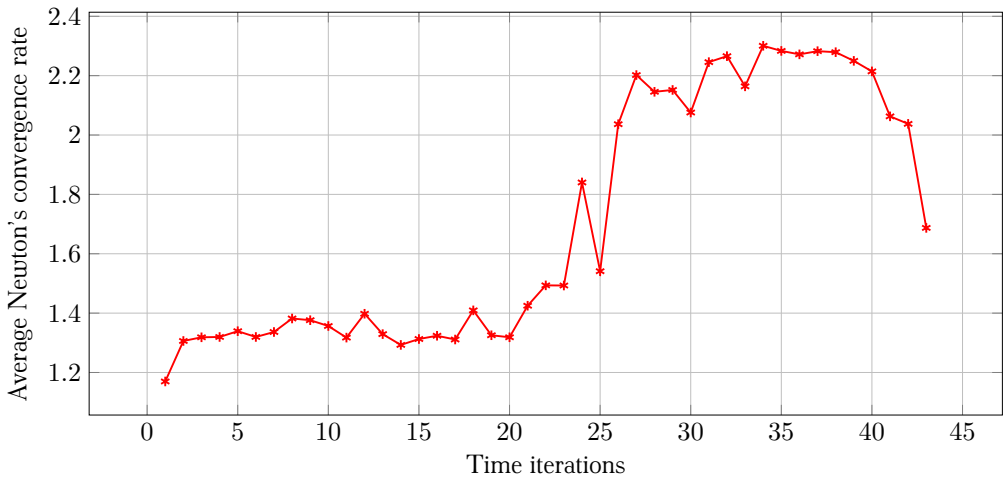


Figure 2.10: Filling test using Kirchhoff transform: evolution of the average Newton's convergence rate during time iterations.

Newton iterations are required. The evolution of the saturation profile during Newton's iterations is reported in Figure 2.11.

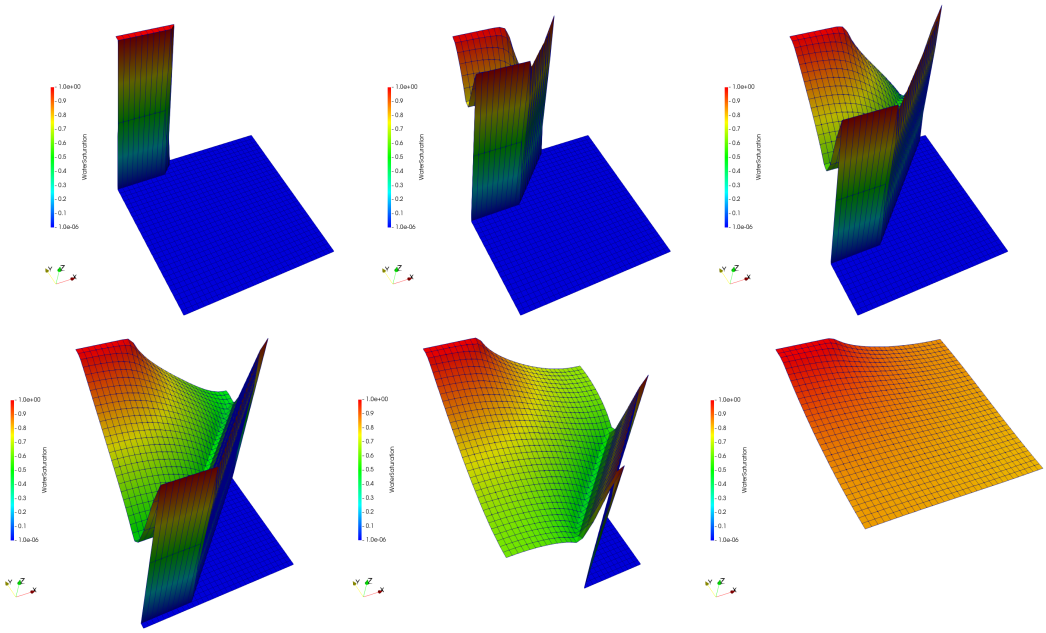


Figure 2.11: Evolution of the saturation profile at Newton's iteration $i = 2, 10, 20, 30, 40, 60$.

Even though one starts from an initial guess $\boldsymbol{\tau}^0$ which is very different from the solution, the Newton method still converges. As expected, the use of a very large time-step introduces an error in the solution. Figure 2.12 shows the distribution over the domain at the final time of the absolute difference between the saturation/Kirchhoff transform values obtained with one single time-step or with several ones. The maximal absolute difference obtained for the saturation is about 10^{-1} and about 10^2 for the Kirchhoff transform.

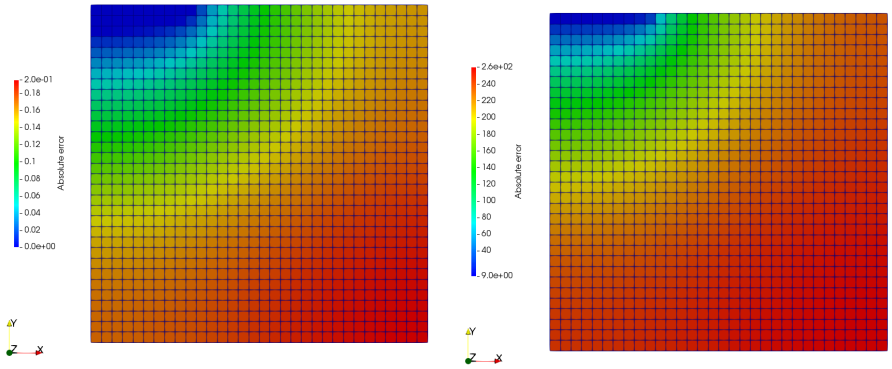


Figure 2.12: Distribution of the absolute error for the saturation values (left) and for the Kirchhoff transform values (right).

2.4.1.3 Draining the domain

In this test, we consider a two-dimensional porous domain $\Omega = [0,1] \times [0,1]$ (in meters), made up of sand which is initially dry in one part of the domain and partially saturated in another one. More precisely we have

$$s_0 = \begin{cases} 1 & \text{in } \Omega_1 = [0, 0.5] \times [0.5, 1], \\ 10^{-6} & \text{in } \Omega_2 = \Omega \backslash \Omega_1. \end{cases}$$

The hydraulic properties of this rock type are the ones given in Table 2.1. During the simulation, water spreads into the domain and leaks out through a portion of the lower boundary $\Gamma^{\mathrm{D}} = \{(x, y) \,|\, x \in [0, 0.3], y = 0\}$ where we impose $u_{\mathrm{D}} = 0$ (which corresponds to a null saturation). A no flux boundary condition is imposed on $\partial\Omega \backslash \Gamma^{\mathrm{D}}$. As for the filling test, for simplicity, the gravity vector is still chosen as $g = -e_y$, $\rho = 1$ and we take water viscosity as $\mu = 10^{-3}\,\mathrm{Pa} \cdot \mathrm{s}$. The configuration of the domain is shown in Figure 2.13. The simulation is performed on a mesh composed of 32×32 cells during a time interval $[0, T]$ with $T = 10^6\,\mathrm{s}$. The time-step is equal to $\Delta t = 2.5 \cdot 10^4\,\mathrm{s}$. The truncation procedure is activated during Newton's iterations.

The values of saturation and Kirchhoff transform obtained at different times are shown in Figure 2.15. Note that this test is very challenging because the Kirchhoff transform varies from values equal to the entry pressure to a null value corresponding to a null saturation. During the simulation we have registered 3 Newton's iterations on average with a maximum of 16 iterations. Figure 2.14 shows the evolution of the average Newton's convergence rate.

Let us now perform the same simulation in just one time-step with $\Delta t = T = 10^6\,\mathrm{s}$. To achieve the convergence, 43 Newton iterations are required. The evolution of the saturation profile during Newton's iteration is reported in Figure 2.16. As expected, the use of a very large time-step introduces an error in the solution. Figure 2.17 shows the distribution over the domain at the final time of the absolute dfference between the saturation/Kirchhoff transform values obtained with one single time-step or with several ones. The maximal absolute difference obtained for the saturation is about 10^{-1} and about 5 for the Kirchhoff transform.

2.4.2 Comparison between different formulations of Richards' equation

In this section, we want to highlight the differences in using the capillary pressure variable (2.2.2)–(2.2.3) or the Kirchhoff transform variable (2.2.7)–(2.2.8), employing the parametrized laws (2.2.1)

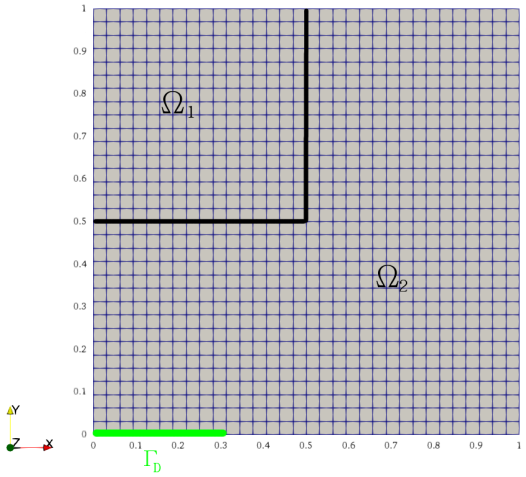


Figure 2.13: Configuration of the domain Ω for the drainage test.

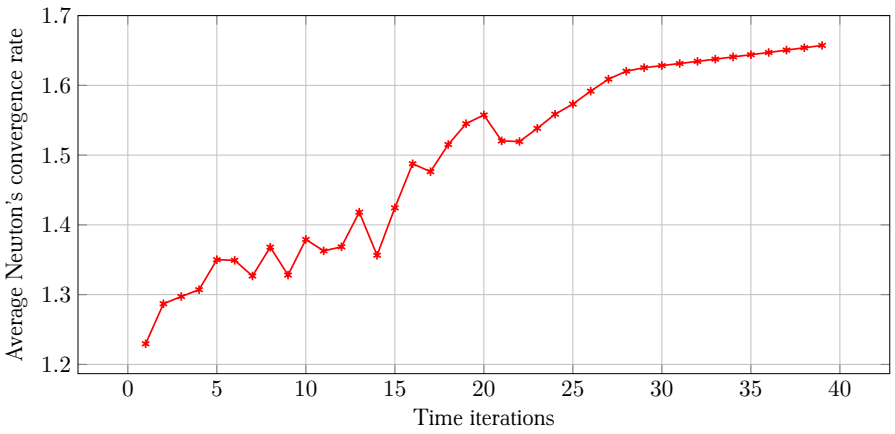


Figure 2.14: Drainage case using Kirchhoff transform: Evolution of the average Newton's convergence rate during time iterations.

and (2.2.6) respectively. As we have already remarked in §1.2.3, the Kirchhoff transform cannot be analytically computed for the van Genuchten-Mualem model entailing a first limit in the use of this variable. Let us perform the solving in pressure formulation the filling test (§2.4.1.2) and drainage test (§2.4.1.3) that we have previously solved using the Kirchhoff transform formulation. In the cited tests, a boundary condition on the Kirchhoff transform variable is stated. For the following test we choose boundary conditions values on water pressure obtained using the Kirchhoff transform equation (1.2.16) in order to replicate faithfully the test settings of the previous section.

2.4.2.1 Filling the domain

Using the test settings presented in §2.4.1.2, we simulate the filling of an initially very dry domain of sand Ω, characterized by $s^0 = 10^{-6}$ as imposed in (2.4.2), in which water is injected through Γ^{D} at pressure $p^{\mathrm{D}} = -3.45761 \cdot 10^3$ Pa $> -p_b = -3.5036 \cdot 10^3$ Pa and a no flux boundary condition is imposed on $\partial\Omega \backslash \Gamma^{\mathrm{D}}$. In Figure 2.5 is shown the domain configuration and in Table 2.1 are reported

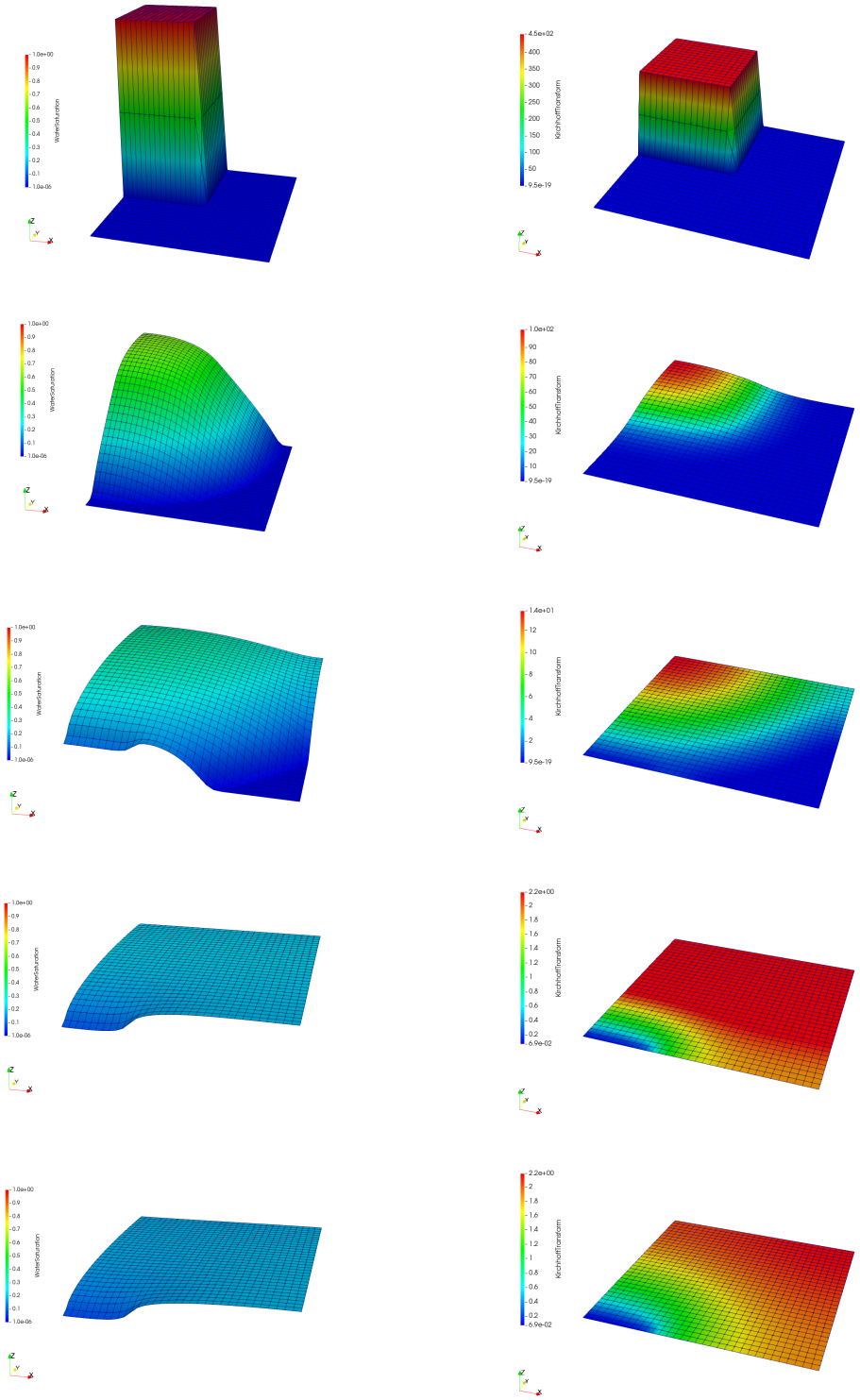


Figure 2.15: Saturation and Kirchhoff transform profiles, on the left and right columns respectively, for $t \in \{0 \cdot \Delta t, \Delta t, 5 \cdot \Delta t, 20 \cdot \Delta t, T\}$.

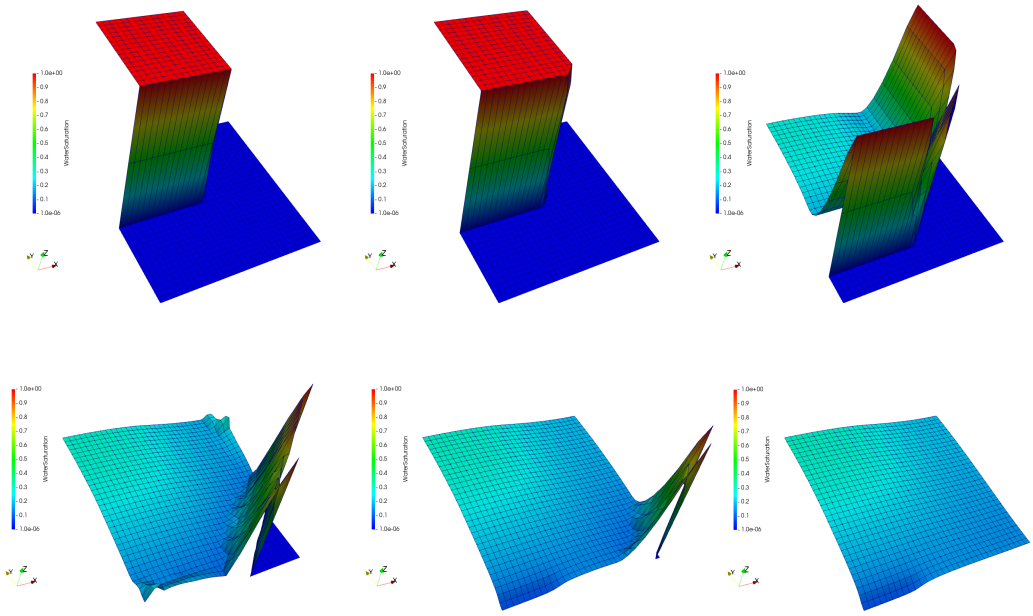


Figure 2.16: Evolution of the saturation profile at nonlinear iteration number $0, 1, 10, 20, 30, 43$.

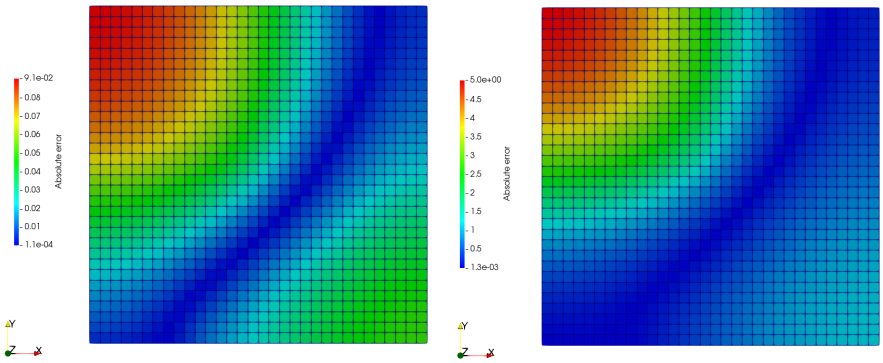


Figure 2.17: Distribution of the absolute error for the saturation values (left) and Kirchhoff transform values (right).

the hydraulic properties of the chosen rock type.

The values of saturation and water pressure obtained at different times are shown in Figure 2.18. During the whole simulation, the Newton method has required 4 iterations on average with a maximum of 46 iterations to converge. Figure 2.19 shows the evolution of the average Newton's convergence rate. The test converges also in case the simulation is carried out using a single step of time ($\Delta t = T$). In this case 103 Newton's iterations are required to converge.

Let us now compare these results with the ones obtained using the Kirchhoff transform formulation in 2.4.1.2. Looking the the saturation profile of both simulations (Figure 2.9 for the Kirchhoff transform formulation and Figure 2.18 for the water pressure one), we can notice that no difference can be remarked. Indeed, we are solving the same problem but using different formulations.

Concerning the computational cost, in Table 2.2 we report data on required Newton's iterations to converge for both formulations and in Figure 2.20 we show the cumulative iteration number evo-

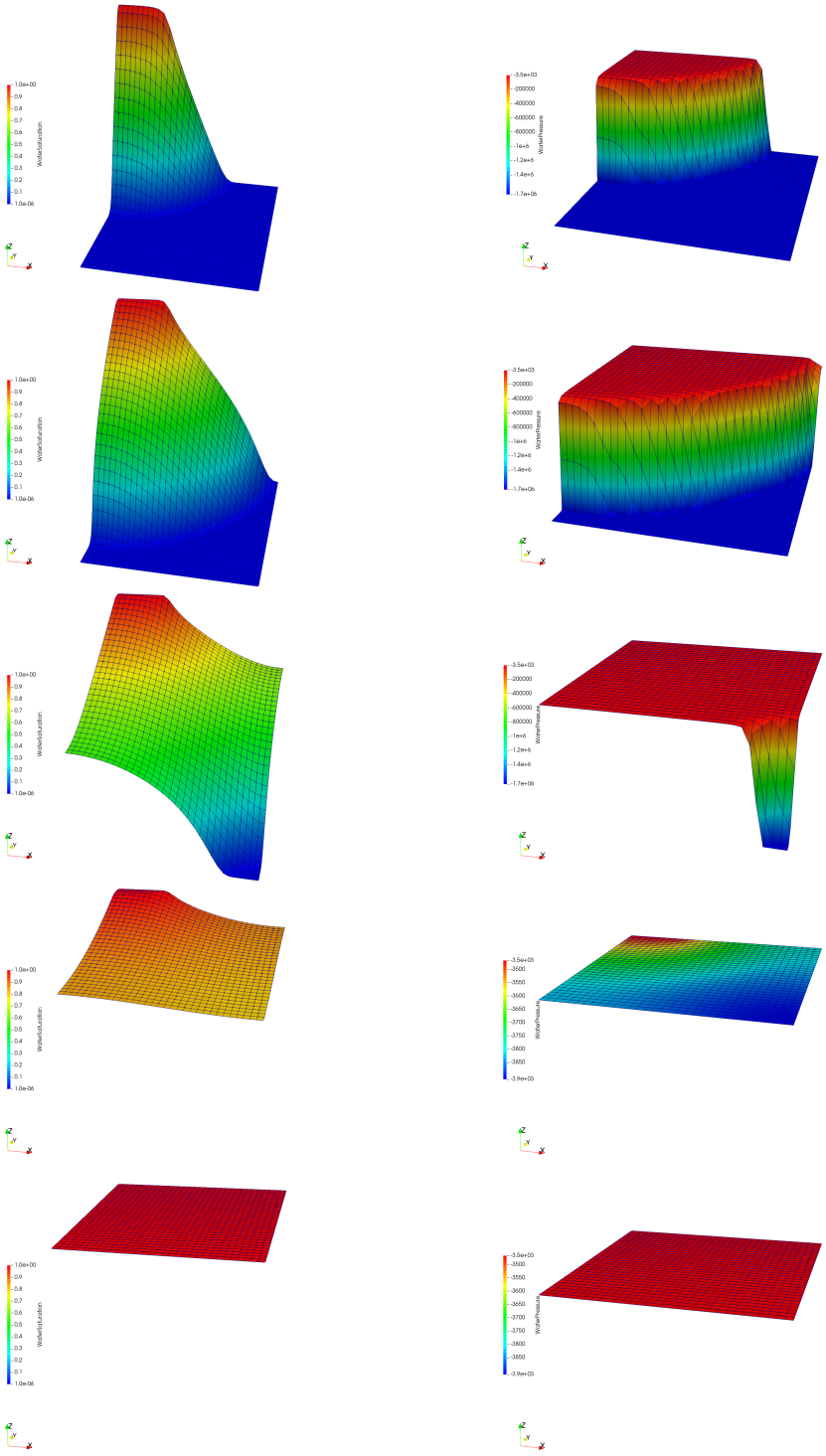


Figure 2.18: Saturation and water pressure profiles, on the left column and right columns respectively, for $t \in \{\Delta t, 5 \cdot \Delta t, 15 \cdot \Delta t, 25 \cdot \Delta t, T\}$

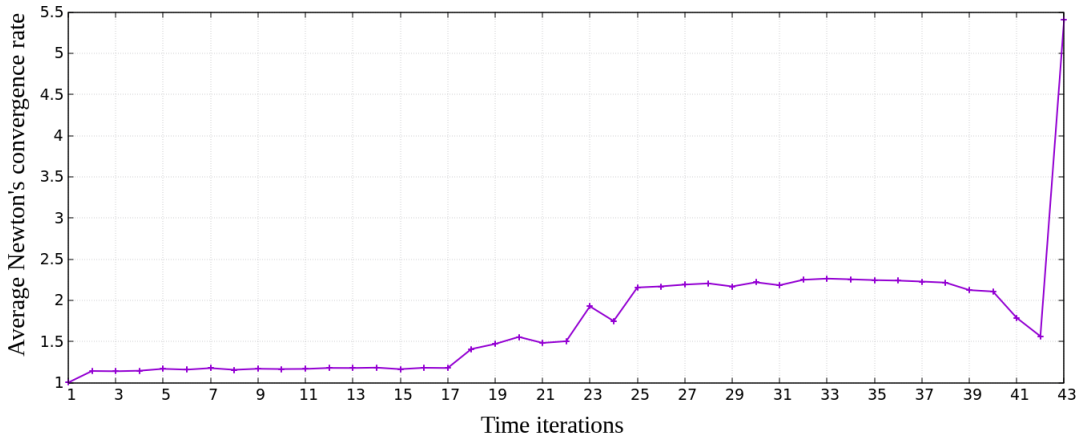


Figure 2.19: Filling test using water pressure: evolution of the average Newton's convergence rate during time iterations.

lution over the whole simulation. We can notice that the evolution profiles of the two formulations follow the same trend. The only difference is that the water pressure formulation requires proportionally more iterations to converge w.r.t. the Kirchhoff transform formulation which benefits of effect of the linearization of the gradient term. Considering the case in which the simulation is performed in one time step, the pressure formulation still required more iterations (almost the double) to converge with respect to the Kirchhoff formulation (103 vs 60 iterations). Finally, observing Figures 2.10 and 2.19, we note that both formulations present almost the same convergence rate trend. More precisely, up to half simulation, Kirchhoff transform formulation has a convergence rate slightly higher than the one of the pressure formulation. Inversely, in the second half of the simulation the pressure formulation has a convergence rate slightly higher than the one characterizing the Kirchhoff transform formulation.

Formulation	# total iterations	# avg iterations	# max iterations
Water pressure	297	4	46
Kirchhoff transform	217	4	17

Table 2.2: Statistics on the required Newton's iterations to converge for filling test using pressure-saturation and Kirchhoff transform-saturation formulation.

2.4.2.2 Draining the domain

Considering the test settings presented in §2.4.1.3, we simulate the draining of the domain Ω, made of sand, which is initially saturated in one part of the domain (§2.4.3), Ω_1, and dry elsewhere, Ω_2. The configuration of the domain is reported in Figure 2.13 and in Table 2.1 are reported the hydraulic properties of the chosen rock type. During the simulation, water radiates into the domain and leaks out through the portion of the lower boundary Γ^{D} where pressure $p^{\mathrm{D}} = -8 \cdot 10^8$ Pa.

The values of saturation and water pressure obtained at different times are shown in Figure 2.21. During the simulation, the Newton method has required 4 iterations on average with a maximum of 44 iterations to converge. Figure 2.22 shows the evolution of the average Newton's convergence rate. This test still converges if we try to solve it in one time iteration ($\Delta t = T$). In this case, the

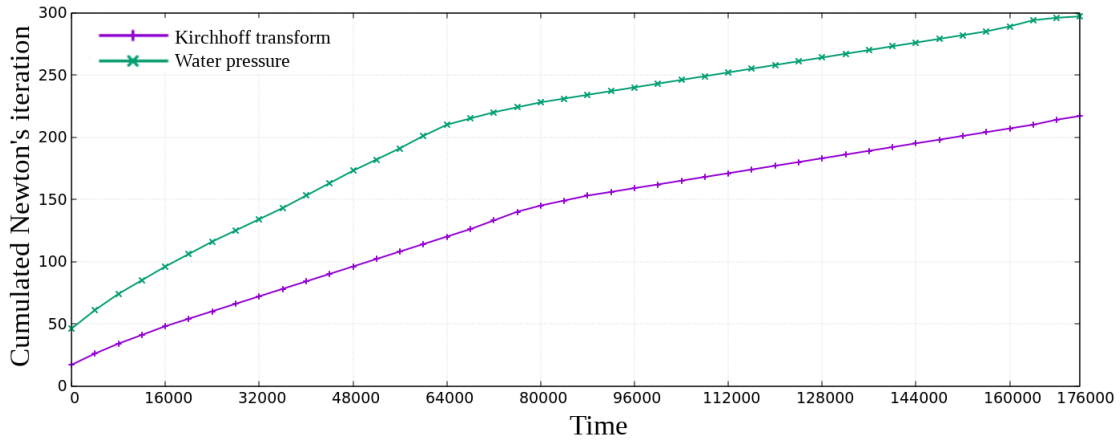


Figure 2.20: Evolution of the Newton's cumulative iterations number required to converge for the filling test case.

Newton method require 65 iterations to converge.

We now compare these results with the ones obtained using the Kirchhoff transform formulation in §2.4.1.3. Scrutinizing Figures 2.15 and 2.21, we can notice that the saturation profile follows the same evolution in both formulations. We can notice that on Γ^{D} and the cells in its neighbourhood, the saturation value is greater for the simulation using the Kirchhoff transform formulation than the value for the simulation in water pressure one. We think this is because the water pressure imposed at the edge is proportionally lower than the value imposed for the Kirchhoff transform. Considering the pressure, the water pressure profile evolution is steeper w.r.t. the one of the Kirchhoff transform.

Regarding the computational cost, we report in Table 2.3 the data on required Newton's iterations to converge for both formulation. Moreover, in Figure 2.23, we show the cumulative iteration number required to converge over the entire simulation. We can notice that its evolution profile has the same trend for both formulation, but the iterations required for the water pressure formulation are always higher than those with Kirchhoff transform formulation, as for the filling test case. Considering the case in which the simulation is performed in one time step, the pressure formulation still required a few more iterations to converge with respect to the Kirchhoff formulation. Finally, in view of Figures 2.14 and 2.22, both formulations present almost the same convergence rate and it evolution follows the same trend. We observe that at the beginning of the simulation, for $t \in [0, 10 \cdot \Delta t]$, the Kirchhoff transform formulation is characterized by a convergence rate slightly higher than the one of the other formulation.

Formulation	# total iterations	# avg iterations	# max iterations
Water pressure	200	4	44
Kirchhoff transform	150	3	16

Table 2.3: Statistics on the required Newton's iterations to converge for drainage test using pressure-saturation and Kirchhoff transform-saturation formulation.

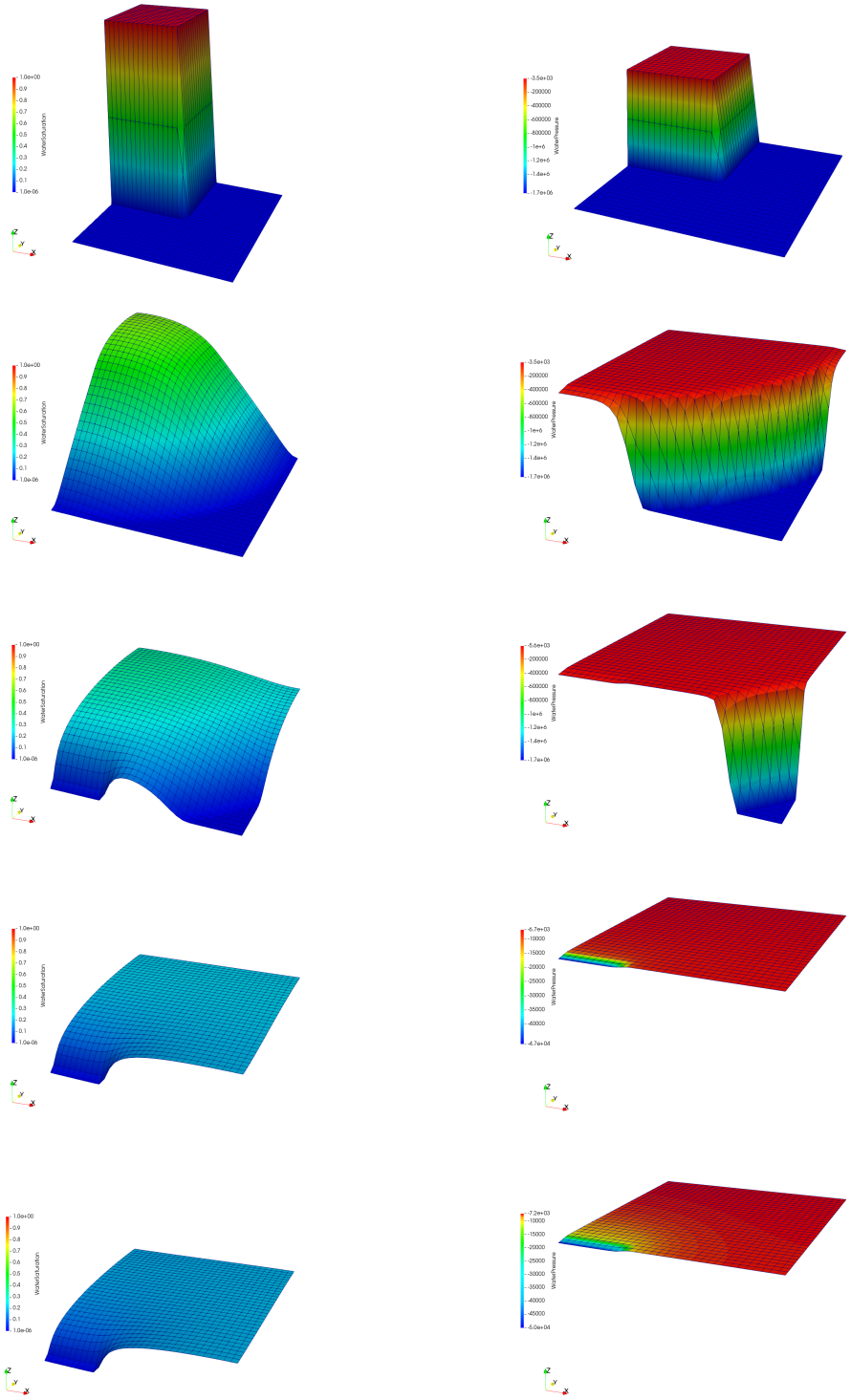


Figure 2.21: Saturation and water pressure profiles, on the left column and right columns respectively, for $t \in \{0 \cdot \Delta t, \Delta t, 5 \cdot \Delta t, 20 \cdot \Delta t, T\}$.

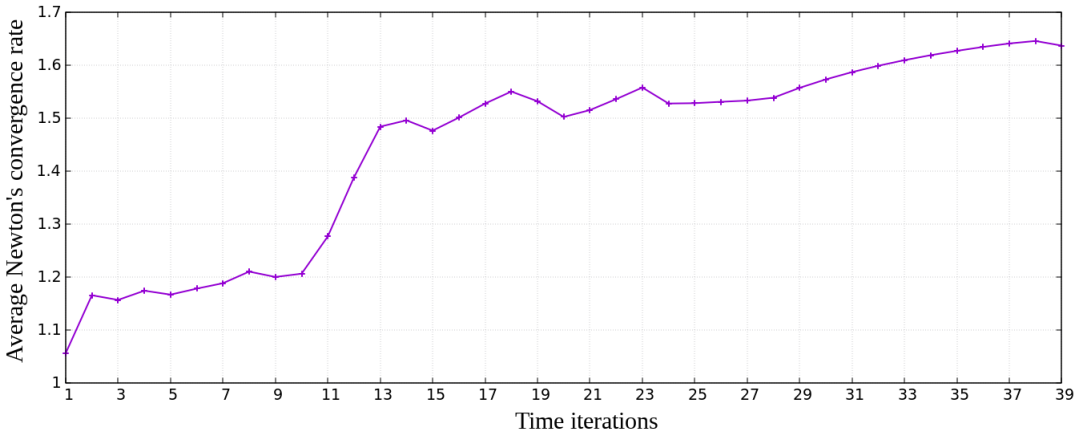


Figure 2.22: Drainage test using water pressure: evolution of the average Newton's convergence rate during time iterations.

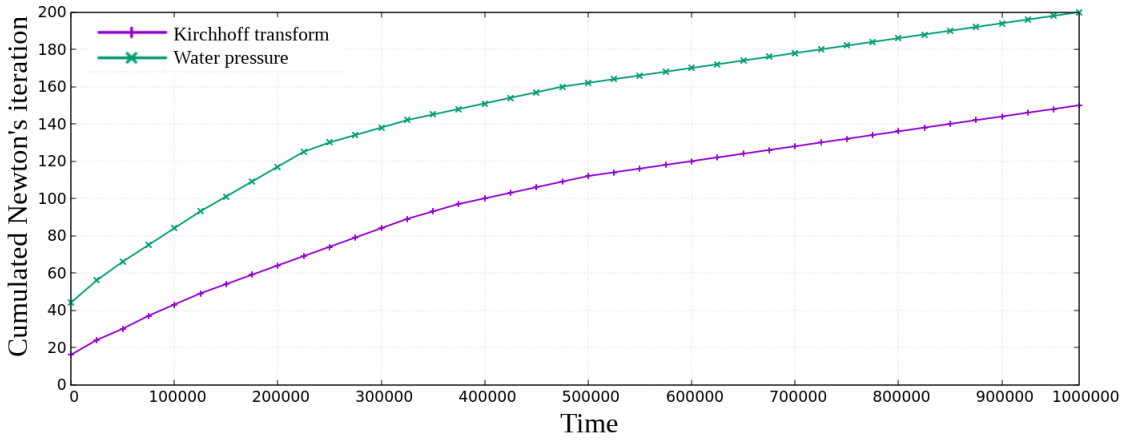


Figure 2.23: Evolution of the Newton's cumulative iterations number required to converge for the drainage test case.

2.4.3 Comparison between different primary variables: saturation, pressure, switching

We wish to show the advantage of using the parametrization technique rather than solving the problem in saturation or pressure variable. Let us recall that, as explained earlier, choosing the pressure as the primary variable is known to be inefficient for dry soils $s \ll 1$; on the other hand, the knowledge of the saturation is not sufficient to describe the pressure in saturated regions where $s = 1 - s_{\mathrm{rn}}$. We now replicate the filling and drainage test cases in water pressure formulation (§2.4.2.1–§2.4.2.2) always considering consider the system (2.2.2)–(2.2.3) and solving it with a fix variable (saturation/pressure) or via a variable switch employing the parametrized laws (2.2.1).

2.4.3.1 Filling the domain

As already said in the remark in §2.4.1.2, the filling test case cannot be solved in saturation variable because the pressure value imposed on the boundary and then, progressively, the pressure value in the whole domain, are greater than the entry pressure vale p_b. We recall that when the saturation reaches value one (so $p \geqslant p_b$), the saturation law is no more invertible. For this reason in saturated regions the saturation variable is not sufficient to describe the pressure profile. On the other hand, solve the problem in pressure variable leads to a non convergence of the Newton method. In Figure 2.24 we report the evolution of the L_∞ norm of the residual at the first time iteration. As expected, the norm of the residual explodes and Newton does not converge. On the other hand, the problem can still be solved using the parametrization technique as previously shown (§2.4.2.1).

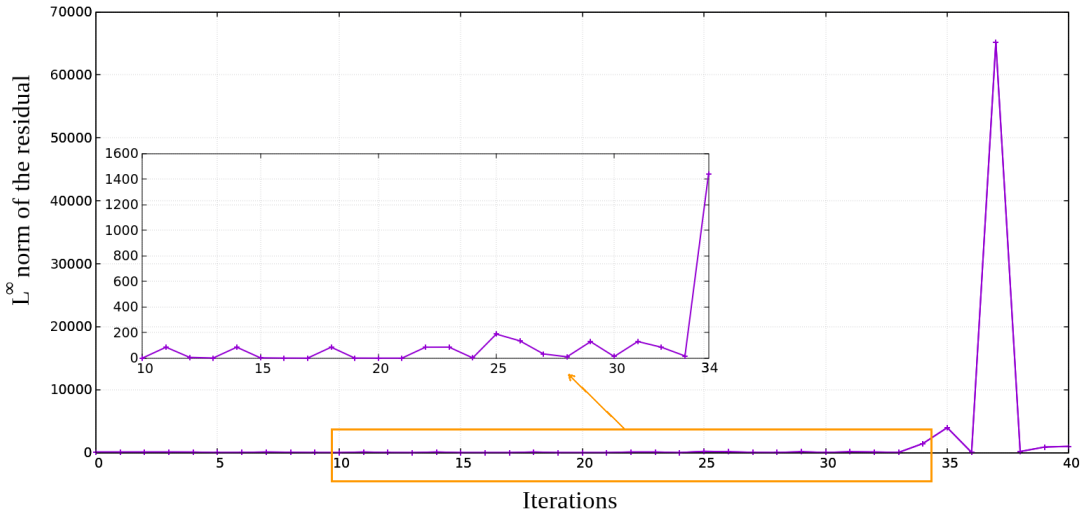


Figure 2.24: Filling test in pressure variable: Evolution of the L^∞ residual norm during the first time step before blow-up.

2.4.3.2 Draining the domain

In the drainage test case, a part of the domain is saturated ($s = 1$ and $p = p_b$) and the rest is dry. Let us recall that the convenient choice as primary unknown to solve the Richards equation in non-saturated region is the saturation variable. This is because if we perform this simulation in pressure variable it fails: the Newton method does not converge because the L^∞ norm of the residual at the first time iteration explodes. We report its evolution in Figure 2.25. On the other

hand, solving the problem in saturation variable is possible because, during the whole simulation, water pressure assumes values smaller ($p = p_b$ only at the initial condition) than the entry pressure.

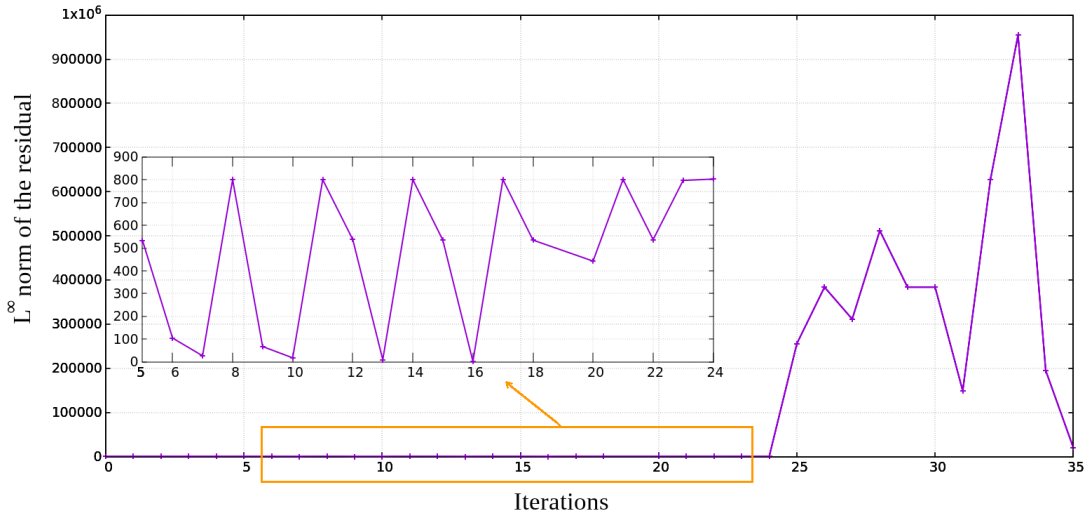


Figure 2.25: Drainage test in pressure variable: Evolution of the L^∞ residual norm during the first time step before blow-up.

Chapter 3

Upstream mobility finite volumes for the Richards equation in heterogeneous domains

The text of this chapter is replicated from [20].

The Richards equation [122] is one of the most well-known simplified models for water filtration in unsaturated soils. While it has been extensively studied in the case of a homogeneous domain, the heterogeneous case seems to have received less attention in the literature, at least from the numerical perspective. The purpose of this chapter is to investigate a class of discretization scheme for a special instance of heterogeneous domains, namely, those with piecewise-uniform physical properties. Our first contribution is to rigorously prove convergence toward a weak solution of cell-centered finite-volume schemes with upstream mobility and without Kirchhoff's transform. Our second contribution is to numerically demonstrate the relevance of locally refining the grid at the interface between subregions, where discontinuities occur, in order to preserve an acceptable accuracy for the results computed with the schemes under consideration.

Before stating our objectives in a precise manner, a few prerequisites must be introduced regarding the model in §3.1–§3.2 and the scheme in §3.4.1–§3.4.2. The goal of this chapter is fully described in §3.3, in relation with other works.

3.1 Richards' equation in heterogeneous porous media

Let $\Omega \subset \mathbb{R}^d$, where $d \in \{2, 3\}$, be a connected open polyhedral domain with Lipschitz boundary $\partial\Omega$. A porous medium defined over the region Ω is characterized by

- the porosity $\phi : \Omega \to (0, 1]$;
- the permeability $\lambda : \Omega \to \mathbb{R}_+^*$;
- the mobility function $\eta : [0, 1] \times \Omega \to \mathbb{R}_+$;
- the saturation law $\mathcal{S} : \mathbb{R} \times \Omega \to [0, 1]$ function of the water pressure and the space location.

The conditions to be satisfied by ϕ, λ, η and $\mathcal{S}$ will be elaborated on later. In a homogeneous medium, these physical properties are uniform over Ω, i.e.,

$$\phi(x) = \phi_0, \qquad \lambda(x) = \lambda_0, \qquad \eta(s, x) = \eta_0(s), \qquad \mathcal{S}(p, x) = \mathcal{S}_0(p)$$

for all $x \in \Omega$. In a heterogeneous medium, the dependence of ϕ, λ, η and $\mathcal{S}$ on x must naturally be taken into account. The quantity s, called saturation, measures the relative volumic presence of

water in the medium. The quantity p is the water pressure, which in our case is the opposite of the capillary pressure.

Let $T > 0$ be a finite time horizon. We designate by $Q_T = (0,T)\times\Omega$ the space-time domain of interest. Our task is to find the saturation field $s : Q_T \to [0,1]$ and the pressure field $p : Q_T \to \mathbb{R}$ so as to satisfy

- the interior equations

$$\phi(x)\,\partial_t s + \nabla\cdot v = 0 \qquad \text{in } Q_T, \tag{3.1.1a}$$
$$v + \lambda(x)\,\eta(s,x)\,\nabla(p - \varrho g\cdot x) = 0 \qquad \text{in } Q_T, \tag{3.1.1b}$$
$$s - \mathcal{S}(p,x) = 0 \qquad \text{in } Q_T; \tag{3.1.1c}$$

- the boundary conditions

$$v\cdot\nu(x) = 0 \qquad \text{on } (0,T)\times\Gamma^{\mathrm{N}}, \tag{3.1.1d}$$
$$p(t,x) = p^{\mathrm{D}}(x) \qquad \text{on } (0,T)\times\Gamma^{\mathrm{D}}; \tag{3.1.1e}$$

- the initial data

$$s(0,x) = s^0(x) \qquad \text{in } \Omega. \tag{3.1.1f}$$

The partial differential equation (3.1.1a) expresses the water volume balance. The flux F involved in this balance is given by the Darcy-Muskat law (3.1.1b), in which g is the gravity vector and ϱ is the known constant density of water, assumed to be incompressible. It is convenient to introduce

$$\psi = -\varrho g\cdot x, \qquad \vartheta = p + \psi, \tag{3.1.2}$$

referred to respectively as gravity potential and hydraulic head. In this way, the Darcy-Muskat law (3.1.1b) can be rewritten as

$$F + \lambda(x)\,\eta(s,x)\,\nabla(p+\psi) = F + \lambda(x)\,\eta(s,x)\,\nabla\vartheta = 0.$$

Equation (3.1.1c) connecting the saturation s and the pressure p is the capillary pressure relation. The boundary $\partial\Omega$ is split into two non-overlapping parts, viz.,

$$\partial\Omega = \Gamma^{\mathrm{N}}\cup\Gamma^{\mathrm{D}}, \qquad \Gamma^{\mathrm{N}}\cap\Gamma^{\mathrm{D}} = \varnothing, \tag{3.1.3}$$

where Γ^{N} is open and Γ^{D} is closed, the latter having a positive $(d-1)$-dimensional Hausdorff measure $m^{d-1}(\Gamma^{\mathrm{D}}) > 0$. The no-flux Neumann condition (3.1.1d) is prescribed on $(0,T)\times\Gamma^{\mathrm{N}}$, where $\nu(x)$ is the outward normal unit vector at $x\in\Gamma^{\mathrm{N}}$. The Dirichlet condition (3.1.1e) with a known Lipschitz function $p^{\mathrm{D}}\in W^{1,\infty}(\Omega)$ is imposed on $(0,T)\times\Gamma^{\mathrm{D}}$. Note that, in our theoretical development, the function p^{D} is assumed to be defined over the whole domain Ω, which is stronger than a data $p^{\mathrm{D}}\in L^\infty(\Gamma^{\mathrm{D}})$ given only on the boundary. The assumption that p^{D} does not depend on time can be removed by following the lines of [44], but we prefer here not to deal with time-dependent boundary data in order to keep the presentation as simple as possible. Finally, the initial data $s^0\in L^\infty(\Omega;[0,1])$ in (3.1.1f) is also a given data.

In this work, we restrict ourselves to a specific type of heterogeneous media, defined as follows. We assume that the domain Ω can be partitioned into several connected polyhedral subdomains

Ω_i, $1 \leqslant i \leqslant I$. Technically, this means that if $\Gamma_{i,j}$ denotes the interface between Ω_i and Ω_j (which can be empty for some particular choices of $\{i,j\}$), then

$$\Omega_i \cap \Omega_j = \varnothing, \quad \overline{\Omega}_i \cap \overline{\Omega}_j = \Gamma_{i,j}, \text{ if } i \neq j, \quad \Omega = \Big(\bigcup_{1 \leqslant i \leqslant I} \Omega_i \Big) \cup \Gamma, \tag{3.1.4}$$

with $\Gamma = \bigcup_{i \neq j} \Gamma_{i,j}$. Each of these subdomains corresponds to a distinctive rocktype. Inside each Ω_i, the physical properties are homogeneous. In other words,

$$\phi(x) = \phi_i, \qquad \lambda(x) = \lambda_i, \qquad \eta(s,x) = \eta_i(s), \qquad \mathcal{S}(p,x) = \mathcal{S}_i(p)$$

for all $x \in \Omega_i$. Therefore, system (3.1.1) is associated with

$$\phi(x) = \sum_{1 \leqslant i \leqslant I} \phi_i \, \mathbf{1}_{\Omega_i}(x), \qquad \eta(s,x) = \sum_{1 \leqslant i \leqslant I} \eta_i(s) \, \mathbf{1}_{\Omega_i}(x), \tag{3.1.5a}$$

$$\lambda(x) = \sum_{1 \leqslant i \leqslant I} \lambda_i \, \mathbf{1}_{\Omega_i}(x), \qquad \mathcal{S}(p,x) = \sum_{1 \leqslant i \leqslant I} \mathcal{S}_i(p) \, \mathbf{1}_{\Omega_i}(x), \tag{3.1.5b}$$

where $\mathbf{1}_{\Omega_i}$ stands for the characteristic function of Ω_i. For all $i \in \{1, \dots, I\}$, we assume that $\phi_i \in (0,1]$ and $\lambda_i > 0$. Furthermore, we require that

$$\eta_i \text{ is increasing on } [0,1], \qquad \eta_i(0) = 0, \qquad \eta_i(1) = \frac{1}{\mu}, \tag{3.1.6a}$$

where $\mu > 0$ is the (known) viscosity of water. In addition to the assumption that $\mathcal{S}(\cdot, x)$, defined in (3.1.5b), is absolutely continuous and nondecreasing, the functions $\mathcal{S}_i$ are also subject to some generic requirements commonly verified the models available in the literature: for each $i \in \{1, \dots, I\}$, there exists $\overline{p}_i \leqslant 0$ such that

$$\mathcal{S}_i \text{ is increasing on } (-\infty, \overline{p}_i], \qquad \lim_{p \to -\infty} \mathcal{S}_i(p) = 0, \qquad \mathcal{S}_i \equiv 1 \text{ on } [\overline{p}_i, +\infty). \tag{3.1.6b}$$

This allows us to define an inverse $\mathcal{S}_i^{-1} : (0,1] \to (-\infty, \overline{p}_i]$ such that $\mathcal{S}_i \circ \mathcal{S}_i^{-1}(s) = s$ for all $s \in (0,1]$. We further assume that for all $i \in \{1, \dots, I\}$ the function $\mathcal{S}_i$ is bounded in $L^1(\mathbb{R}_-)$, or equivalently, that $\mathcal{S}_i^{-1} \in L^1(0,1)$. It thus makes sense to consider the capillary energy density functions $\mathfrak{E}_i : \mathbb{R} \times \Omega_i \to \mathbb{R}_+$ defined by

$$\mathfrak{E}_i(s,x) = \int_{\mathcal{S}_i(p^{\mathrm{D}}(x))}^{s} \phi_i(\mathcal{S}_i^{-1}(\varsigma) - p^{\mathrm{D}}(x)) \, \mathrm{d}\varsigma. \tag{3.1.7}$$

For all $x \in \Omega_i$, the function $\mathfrak{E}_i(\cdot, x)$ is nonnegative, convex since $\mathcal{S}_i^{-1}$ is monotone, and bounded on $[0,1]$ as a consequence of the integrability of $\mathcal{S}_i$. For technical reasons that will appear clearly later on, we further assume that

$$\sqrt{\eta_i \circ \mathcal{S}_i} \in L^1(\mathbb{R}_-), \qquad \forall i \in \{1, \dots, I\}. \tag{3.1.8}$$

Let $Q_{i,T} = (0,T) \times \Omega_i$ be the space-time subdomains for $1 \leqslant i \leqslant I$. The interior equations (3.1.1a)–(3.1.1c) then boil down to

$$\phi_i \, \partial_t s + \nabla \cdot v = 0 \qquad \text{in } Q_{i,T}, \tag{3.1.9a}$$

$$v + \lambda_i \, \eta_i \, \nabla(p + \psi) = 0 \qquad \text{in } Q_{i,T}, \tag{3.1.9b}$$

$$s - \mathcal{S}_i(p) = 0 \qquad \text{in } Q_{i,T}. \tag{3.1.9c}$$

At the interface $\Gamma_{i,j}$ between Ω_i and Ω_j, $i \neq j$, any solution of (3.1.1a)–(3.1.1c) satisfies the matching conditions

$$v_i \cdot \nu_i + v_j \cdot \nu_j = 0 \qquad \text{on } (0,T) \times \Gamma_{i,j}, \tag{3.1.10a}$$

$$p_i - p_j = 0 \qquad \text{on } (0,T) \times \Gamma_{i,j}. \tag{3.1.10b}$$

In the continuity of the normal fluxes (3.1.10a), which is enforced by the conservation of water volume, ν_i denotes the outward normal to $\partial\Omega_i$ and $F_i \cdot \nu_i$ stands for the trace of the normal component of $F_{|_{Q_{i,T}}}$ on $(0,T) \times \partial\Omega_i$. In the continuity of pressure (3.1.10b), which also results from (3.1.1a)–(3.1.1c), p_i denotes the trace on $(0,T) \times \partial\Omega_i$ of the pressure $p_{|_{Q_{i,T}}}$ in the i-th domain.

3.2 Stability features and notion of weak solutions

We wish to give a proper sense to the notion of weak solution for problem (3.1.1). To achieve this purpose, we need a few mathematical transformations the definition of which crucially relies on a fundamental energy estimate at the continuous level. The calculations below are aimed at highlighting this energy estimate and will be carried out in a formal way, in constrast to those in the fully discrete setting.

Multiplying (3.1.9a) by $p - p^{\mathrm{D}}$, invoking (3.1.7), integrating over Ω_i and summing over i, we end up with

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{I} \int_{\Omega_i} €_i(s,x)\,\mathrm{d}x + \sum_{i=1}^{I} \int_{\Omega_i} (\nabla \cdot v)(p - p^{\mathrm{D}})\,\mathrm{d}x = 0. \tag{3.2.1}$$

We now integrate by parts the second term. Thanks to the matching conditions (3.1.10) and the regularity of p^{D}, we obtain

$$\mathtt{A} := \sum_{i=1}^{I} \int_{\Omega_i} (\nabla \cdot v)(p - p^{\mathrm{D}})\,\mathrm{d}x = -\sum_{i=1}^{I} \int_{\Omega_i} v \cdot \nabla(p - p^{\mathrm{D}})\,\mathrm{d}x.$$

It follows from the flux value (3.1.9b) that

$$\begin{aligned}
\mathtt{A} &= \sum_{i=1}^{I} \int_{\Omega_i} \lambda_i \eta_i(s) \nabla(p + \psi) \cdot \nabla(p - p^{\mathrm{D}})\,\mathrm{d}x \\
&= \sum_{i=1}^{I} \int_{\Omega_i} \lambda_i \eta_i(s) |\nabla p|^2\,\mathrm{d}x - \sum_{i=1}^{I} \int_{\Omega_i} \lambda_i \eta_i(s) \nabla\psi \cdot \nabla p^{\mathrm{D}}\,\mathrm{d}x \\
&+ \sum_{i=1}^{I} \int_{\Omega_i} \lambda_i \eta_i(s) \nabla p \cdot \nabla(\psi - p^{\mathrm{D}})\,\mathrm{d}x.
\end{aligned}$$

Young's inequality, combined with the boundedness of ∇p^{D}, $\nabla\psi$, λ and η, yields

$$\mathtt{A} \geqslant \frac{1}{2} \sum_{i=1}^{I} \int_{\Omega_i} \lambda_i \eta_i(s) |\nabla p|^2\,\mathrm{d}x - C$$

for some $C \geqslant 0$ depending only on λ, η, ψ, μ, Ω and p^{D}.

Let us define the energy $\mathfrak{E} : [0,T] \to \mathbb{R}_+$ by

$$\mathfrak{E}(t) = \sum_{i=1}^{I} \int_{\Omega_i} \mathfrak{E}_i(s(t,x),x)\,\mathrm{d}x, \qquad 0 \leqslant t \leqslant T.$$

Integrating (3.2.1) w.r.t. time results in

$$\mathfrak{E}(T) + \frac{1}{2}\sum_{i=1}^{I} \iint_{Q_{i,T}} \lambda_i \eta_i(s) |\nabla p|^2 \,\mathrm{d}x\,\mathrm{d}t \leqslant \mathfrak{E}(0) + CT. \tag{3.2.2}$$

Estimate (3.2.2) is the core of our analysis. However, it is difficult to use in its present form since $\eta_i(s) = \eta_i(\mathcal{S}_i(p))$ vanishes as p tends to $-\infty$, so that the control of ∇p degenerates. To circumvent this difficulty, we resort to the nonlinear functions (customarily referred to as the Kirchhoff transforms) $\Theta_i : \mathbb{R} \to \mathbb{R}$, $\Phi_i : \mathbb{R} \to \mathbb{R}$, and $\Upsilon : \mathbb{R} \times \Omega \to \mathbb{R}$ respectively defined by

$$\Theta_i(p) = \int_0^p \sqrt{\lambda_i \eta_i \circ \mathcal{S}_i(\pi)}\,\mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{3.2.3a}$$

$$\Phi_i(p) = \int_0^p \lambda_i \eta_i \circ \mathcal{S}_i(\pi)\,\mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{3.2.3b}$$

$$\Upsilon(p) = \int_0^p \min_{1\leqslant i\leqslant I} \sqrt{\lambda_i \eta_i \circ \mathcal{S}_i(\pi)}\,\mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{3.2.3c}$$

the notion of Υ being due to [65]. Bearing in mind that $\mathfrak{E}(T) \geqslant 0$, estimate (3.2.2) implies that

$$\sum_{i=1}^{I} \iint_{Q_{i,T}} |\nabla \Theta_i(p)|^2 \,\mathrm{d}x\,\mathrm{d}t \leqslant 2(\mathfrak{E}(0) + CT) < +\infty. \tag{3.2.4}$$

As $\Phi_i \circ \Theta_i^{-1}$ is Lipschitz continuous, this also gives rise to a $L^2(Q_{i,T})$-estimate on $\nabla \Phi_i(p)$. The functions $\sum_i \Theta_i(p)\mathbf{1}_{\Omega_i}$ and $\sum_i \Phi_i(p)\mathbf{1}_{\Omega_i}$ are in general discontinuous across the interfaces $\Gamma_{i,j}$, unlike $\Upsilon(p)$. Since the functions $\Upsilon \circ \Theta_i^{-1}$ are Lipschitz continuous, we can readily infer from (3.2.4) that

$$\iint_{Q_T} |\nabla \Upsilon(p)|^2 \,\mathrm{d}x \leqslant C \tag{3.2.5}$$

for some C depending on T, Ω, $\|\nabla p^{\mathrm{D}}\|_\infty$, the $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$'s and

$$\overline{\lambda} = \|\lambda\|_{L^\infty(\Omega)} = \max_{1\leqslant i\leqslant I} \lambda_i, \qquad \overline{\eta} = \|\eta\|_{L^\infty(\Omega)} = \max_{1\leqslant i\leqslant I} \|\eta_i\|_{L^\infty(\Omega)} = \frac{1}{\mu},$$

the last equality being due to (3.1.6a).

Moreover, $\Upsilon(p) - \Upsilon(p^{\mathrm{D}})$ vanishes on $(0,T)\times\Gamma^{\mathrm{D}}$. Poincaré's inequality provides a $L^2(Q_T)$-estimate on $\Upsilon(p)$ since Γ^{D} has positive measure and since $\Upsilon(p^{\mathrm{D}})$ is bounded in Ω. In view of assumption (3.1.8), the functions Θ_i and Υ are bounded on $\mathbb{R}_-$. Besides, for $p \geqslant 0$, $\eta_i \circ \mathcal{S}_i(p) = 1/\mu$, so that $\Theta_i(p) = p\sqrt{\lambda_i/\mu}$ and $\Upsilon(p) = \min_{1\leqslant i\leqslant I} p\sqrt{\lambda_i/\mu}$. It finally comes that

$$\Theta_i(p) \leqslant C(1 + \Upsilon(p)), \qquad \forall p \in \mathbb{R},\ 1 \leqslant i \leqslant I, \tag{3.2.6}$$

from which we infer a $L^2(Q_{i,T})$-estimate on $\Theta_i(p)$. Putting

$$V = \big\{u \in H^1(\Omega) \mid u_{|_{\Gamma^{\mathrm{D}}}} = 0\big\},$$

the above estimates suggest the following notion of weak solution for our problem.

Definition 3.2.1. *A measurable function $p : Q_T \to \mathbb{R}$ is said to be a weak solution to the problem* (3.1.9a)–(3.1.9c) *if*

$$\Theta_i(p) \in L^2((0,T);H^1(\Omega_i)), \qquad \text{for } 1 \leqslant i \leqslant I, \tag{3.2.7a}$$

$$\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T);V) \tag{3.2.7b}$$

and if for all $\varphi \in C_c^\infty([0,T) \times (\Omega \cup \Gamma^{\mathrm{N}}))$, there holds

$$\iint_{Q_T} \phi\, \mathcal{S}(p,x)\partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t + \int_\Omega \phi\, s^0\varphi(\cdot,0)\,\mathrm{d}x + \iint_{Q_T} F\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t = 0, \tag{3.2.7c}$$

with

$$v = -\nabla\Phi_i(p) + \lambda_i\eta_i(\mathcal{S}_i(p))\,\varrho g \qquad \text{in } Q_{i,T},\ 1 \leqslant i \leqslant I. \tag{3.2.7d}$$

The expression (3.2.7d) is a reformulation of the original one (3.1.9b) in a quasilinear form which is suitable for analysis, even though the physical meaning of the Kirchhoff transform $\Phi_i(p)$ is unclear. While the formulation (3.2.7c) should be thought of as a weak form of (3.1.9a), (3.1.10a), (3.1.1f), and (3.1.1d), the condition $\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T);V)$ contains (3.1.10b) and (3.1.1e).

3.3 Goal and positioning of this chapter

We are now in a position to clearly state the two objectives of this chapter.

The first objective is to put forward a rigorous proof that, for problem (3.1.1) with heterogeneous data (3.1.5), cell-centered finite-volume schemes with upstream mobility such as described in §3.4.2, do converge towards a weak solution (in the sense of Definition 3.2.1) as the discretization parameters tend to 0. Such mathematically assessed convergence results are often dedicated to homogeneous cases: see for instance [16, 75, 120] for schemes involving the Kirchhoff transforms for Richards' equation, [5] for a upstream mobility CVFE approximation of Richards' equation in anisotropic domains, [51, 53, 54] for schemes for two-phase flows involving the Kirchhoff transform, and [76, 85] for upstream mobility schemes for two-phase porous media flows. For flows in highly heterogeneous porous media, rigorous mathematical results have been obtained for schemes involving the introduction of additional interface unknowns and Kirchhoff's transforms (see for instance [30, 40, 41, 65]), or under the non-physical assumption that the mobilities are strictly positive [72, 74]. We also refer the reader to [15, 121] where the assumption of the non-degeneracy of the mobility has been made. It was established very recently in [32] that cell-centered finite-volumes with (hybrid) upwinding also converge for two-phase flows in heterogeneous domains, but with a specific treatment of the interfaces located at the heterogeneities. Here, the novelty lies in the fact that we do not consider any specific treatment of the interface in the design of the scheme.

3.4 Finite-volume discretization

The scheme we consider in this chapter is based on two-point flux approximation (TPFA) finite-volumes. Hence, it is subject to some restrictions on the mesh [71, 82]. We first review the requirements on the mesh in §3.4.1. Next, we construct the upstream mobility finite-volume scheme for Richards' equation in §3.4.2. The main mathematical results of this chapter, which are the well-posedness of the nonlinear system corresponding to the scheme and the convergence of the scheme, are then summarized in §3.4.3.

3.4.1 Admissible discretization of Q_T

Let us start by discretizing w.r.t. space.

Definition 3.4.1. *An* admissible mesh of Ω *is a triplet* $(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}})$ *such that the following conditions are fulfilled:*

(i) Each control volume (or cell) $K \in \mathscr{T}$ *is non-empty, open, polyhedral and convex, with positive* d*-dimensional Lebesgue measure* $m_K > 0$*. We assume that*

$$K \cap L = \varnothing \quad \text{if } K, L \in \mathscr{T} \text{ with } K \neq L, \qquad \text{while} \quad \bigcup_{K\in\mathscr{T}} \overline{K} = \overline{\Omega}.$$

Moreover, we assume that the mesh is adapted to the heterogeneities of Ω*, in the sense that for all* $K \in \mathscr{T}$*, there exists* $i \in \{1, \dots, I\}$ *such that* $K \subset \Omega_i$*.*

(ii) Each face $\sigma \in \mathscr{E}$ *is closed and is contained in a hyperplane of* $\mathbb{R}^d$*, with positive* $(d-1)$*-dimensional Hausdorff measure* $m^{d-1}(\sigma) = m_\sigma > 0$*. We assume that* $m^{d-1}(\sigma \cap \sigma') = 0$ *for* $\sigma, \sigma' \in \mathscr{E}$ *unless* $\sigma' = \sigma$*. For all* $K \in \mathscr{T}$*, we assume that there exists a subset* $\mathscr{E}_K$ *of* $\mathscr{E}$ *such that* $\partial K = \bigcup_{\sigma\in\mathscr{E}_K} \sigma$*. Moreover, we suppose that* $\bigcup_{K\in\mathscr{T}} \mathscr{E}_K = \mathscr{E}$*. Given two distinct control volumes* $K, L \in \mathscr{T}$*, the intersection* $\overline{K} \cap \overline{L}$ *either reduces to a single face* $\sigma \in \mathscr{E}$ *denoted by* $K|L$*, or its* $(d-1)$*-dimensional Hausdorff measure is* 0*.*

(iii) The cell-centers $(x_K)_{K\in\mathscr{T}}$ *are pairwise distinct with* $x_K \in K$*, and are such that, if* $K, L \in \mathscr{T}$ *share a face* $K|L$*, then the vector* $x_L - x_K$ *is orthogonal to* $K|L$*.*

(iv) For the boundary faces $\sigma \subset \partial\Omega$*, we assume that either* $\sigma \subset \Gamma^{\mathrm{D}}$ *or* $\sigma \subset \overline{\Gamma}^{\mathrm{N}}$*. For* $\sigma \subset \partial\Omega$ *with* $\sigma \in \mathscr{E}_K$ *for some* $K \in \mathscr{T}$*, we assume additionally that there exists* $x_\sigma \in \sigma$ *such that* $x_\sigma - x_K$ *is orthogonal to* σ*.*

In our problem, the standard Definition 3.4.1 must be supplemented by a compatibility property between the mesh and the subdomains. By "compatbility" we mean that each cell must lie entirely inside a single subregion. Put another way,

$$\forall K \in \mathscr{T}, \quad \exists!\, i(K) \in \{1, \dots, I\} \;\mid\; K \subset \Omega_{i(K)}. \tag{3.4.1}$$

This has two consequences. The first one is that, if we define

$$\mathscr{T}_i = \{K \in \mathscr{T} \mid K \subset \Omega_i\}, \qquad 1 \leqslant i \leqslant I, \tag{3.4.2}$$

then $\mathscr{T} = \bigcup_{i=1}^{I} \mathscr{T}_i$. The second one is that the subdomain interfaces $\Gamma_{i,j}$ for $i \neq j$ coincide necessarily with some edges $\sigma \in \mathscr{E}$. To express this more accurately, let $\mathscr{E}_\Gamma = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma\}$ be the set of the interface edges, $\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma^{\mathrm{D}}\}$ be the set of Dirichlet boundary edges, and $\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}} = \{\sigma \in \mathscr{E} \mid \sigma \subset \overline{\Gamma}^{\mathrm{N}}\}$ be the set of Neumann boundary edges. Then, $\Gamma = \bigcup_{\sigma\in\mathscr{E}_\Gamma} \sigma$, while $\Gamma^{\mathrm{D}} = \bigcup_{\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}} \sigma$ and $\overline{\Gamma}^{\mathrm{N}} = \bigcup_{\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}} \sigma$. For later use, it is also convenient to introduce the subset $\mathscr{E}_i \subset \mathscr{E}$ consisting of those edges that correspond to cells in $\mathscr{T}_i$ only, i.e.,

$$\mathscr{E}_i = \Big(\bigcup_{K\in\mathscr{T}_i} \mathscr{E}_K \Big) \backslash \mathscr{E}_\Gamma, \qquad 1 \leqslant i \leqslant I, \tag{3.4.3a}$$

and the subset $\mathscr{E}_{\mathrm{int}}$ of the internal edges, i.e.,

$$\mathscr{E}_{\mathrm{int}} = \mathscr{E} \backslash (\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}) = \bigcup_{K,L\in\mathscr{T}} \{\sigma = K|L\}. \tag{3.4.3b}$$

Note that $\mathscr{E}_\Gamma \subset \mathscr{E}_{\rm int}$.

To each edge $\sigma \in \mathscr{E}$, we associate a distance d_σ by setting

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathscr{E}_{\rm int}, \\ |x_K - x_\sigma| & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\rm ext}^{\rm D} \cup \mathscr{E}_{\rm ext}^{\rm N}). \end{cases} \tag{3.4.4}$$

We also define $d_{K\sigma} = \mathrm{dist}(x_K, \sigma)$ for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$. The transmissivity of the edge $\sigma \in \mathscr{E}$ is defined by

$$a_\sigma = \frac{m_\sigma}{d_\sigma}. \tag{3.4.5}$$

Throughout this chapter, many discrete quantities $\boldsymbol{u}$ will be defined either in cells $K \in \mathscr{T}$ or on Dirichlet boundary edges $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$, i.e. $\boldsymbol{w} = ((w_K)_{K\in\mathscr{T}}, (w_\sigma)_{\sigma\in\mathscr{E}_{\rm ext}^{\rm D}}) \in \mathbb{X}^{\mathscr{T}\cup\mathscr{E}_{\rm ext}^{\rm D}}$, where $\mathbb{X}$ can be either $\mathbb{R}^\ell$, $\ell \geqslant 1$, or a space of functions. Then for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$, we define the mirror value $w_{K\sigma}$ by

$$w_{K\sigma} = \begin{cases} w_L & \text{if } \sigma = K|L \in \mathscr{E}_{\rm int}, \\ w_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \\ w_\sigma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm D}. \end{cases} \tag{3.4.6}$$

The diamond cell Δ_σ corresponding to the edge σ is defined as the convex hull of $\{x_K, x_{K\sigma}, \sigma\}$ for K such that $\sigma \in \mathscr{E}_K$, while the half-diamond cell $\Delta_{K\sigma}$ is defined as the convex hull of $\{x_K, \sigma\}$. Denoting by m_{Δ_σ} the Lebesgue measure of Δ_σ, the elementary geometrical relation $m_{\Delta_\sigma} = d\, m_\sigma d_\sigma$ where d stands for the dimension will be used many times in what follows.

Another notational shorthand is worth introducing now, since it will come in handy in the sequel. Let

$$f(\cdot, x) = \sum_{1\leqslant i\leqslant I} f_i(\cdot)\mathbf{1}_{\Omega_i}(x) \tag{3.4.7a}$$

be a scalar quantity or a function whose dependence of $x \in \Omega$ is of the type (3.1.5). Then, for $K \in \mathscr{T}$, we slightly abuse the notations in writing

$$f_K(\cdot) := f(\cdot, x_K) = f_{i(K)}(\cdot), \tag{3.4.7b}$$

where the index $i(K)$ is defined in (3.4.1). The last equality in the above equation holds by virtue of the compatibility property. For example, we will have not only $\phi_K = \phi(x_K)$, $\lambda_K = \lambda(x_K)$, $\eta_K(s) = \eta(s, x_K)$, $\mathcal{S}_K(p) = \mathcal{S}(p, x_K)$ but also $\mathbb{C}_K(s) = \mathbb{C}(s, x_K)$. Likewise, we shall be writing $f_{K\sigma}(\cdot) = f(\cdot, x_{K\sigma})$ for the mirror cell without any ambiguity: if $\sigma \in \mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm N}$, then $x_{K\sigma}$ is a cell-center; if $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$, then $x_{K\sigma}$ lies on the boundary but does not belong to an interface between subdomains.

The size $h_{\mathscr{T}}$ and the regularity $\zeta_{\mathscr{T}}$ of the mesh are respectively defined by

$$h_{\mathscr{T}} = \max_{K\in\mathscr{T}} \mathrm{diam}(K), \qquad \zeta_{\mathscr{T}} = \min_{K\in\mathscr{T}} \left(\frac{1}{\mathrm{Card}\ \mathscr{E}_K} \min_{\sigma\in\mathscr{E}_K} \frac{d_{K\sigma}}{\mathrm{diam}(K)} \right). \tag{3.4.8}$$

The time discretization is given by $(t^n)_{0\leqslant 1\leqslant N}$ with $0 = t^0 < t^1 < \cdots < t^N = T$. We denote by $\Delta t^n = t^n - t^{n-1}$ for all $n \in \{1, \ldots, N\}$ and by $\boldsymbol{\Delta t} = (\Delta t^n)_{1\leqslant n\leqslant N}$.

3.4.2 Upstream mobility TPFA Finite Volume scheme

Given a discrete saturation profile $(s_K^{n-1})_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$ at time t^{n-1}, $n \in \{1,\dots,N\}$, we seek for a discrete pressure profile $(p_K^n)_{K\in\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ at time t^n solution to the following nonlinear system of equations. Taking advantage of the notational shorthand (3.4.7b), we define

$$s_K^n = \mathcal{S}_K(p_K^n), \qquad K \in \mathscr{T},\ n \geqslant 1. \tag{3.4.9}$$

The volume balance (3.1.9a) is then discretized into

$$m_K\phi_K\frac{s_K^n - s_K^{n-1}}{\Delta t^n} + \sum_{\sigma\in\mathscr{E}_K} F_{K\sigma}^n = 0, \qquad K \in \mathscr{T},\ n \geqslant 1, \tag{3.4.10}$$

using the approximation

$$F_{K\sigma}^n = \frac{m_\sigma}{d_\sigma}\lambda_\sigma\eta_\sigma^n(\vartheta_K^n - \vartheta_{K\sigma}^n), \quad \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ n \geqslant 1, \tag{3.4.11a}$$

for the flux (3.1.1b), with

$$\vartheta_K^n = p_K^n + \psi_K, \qquad \vartheta_{K\sigma}^n = p_{K\sigma}^n + \psi_{K\sigma}, \tag{3.4.11b}$$

where the mirror values $p_{K\sigma}^n$ and $\psi_{K\sigma}$ are given by (3.4.6). In the numerical flux (3.4.11a), the edge permeabilities $(\lambda_\sigma)_{\sigma\in\mathscr{E}}$ are set to

$$\lambda_\sigma = \begin{cases} \dfrac{\lambda_K\lambda_L d_\sigma}{\lambda_K d_{L,\sigma} + \lambda_L d_{K,\sigma}} & \text{if } \sigma = K|L \in \mathscr{E}_{\rm int},\\ \lambda_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}, \end{cases}$$

while the edge mobilities are upwinded according to

$$\eta_\sigma^n = \begin{cases} \eta_K(s_K^n) & \text{if } \vartheta_K^n > \vartheta_{K\sigma}^n,\\ \frac{1}{2}(\eta_K(s_K^n) + \eta_{K\sigma}(s_{K\sigma}^n)) & \text{if } \vartheta_K^n = \vartheta_{K\sigma}^n,\\ \eta_{K\sigma}(s_{K\sigma}^n) & \text{if } \vartheta_K^n < \vartheta_{K\sigma}^n. \end{cases} \tag{3.4.11c}$$

In practice, the definition of η_σ^n when $\vartheta_K^n = \vartheta_{K\sigma}^n$ has no influence on the scheme. We choose here to give a symmetric definition that does not depend on the orientation of the edge σ in order to avoid ambiguities.

The boundary condition $p^{\rm D}$ is discretized into

$$\begin{cases} p_K^{\rm D} = \frac{1}{m_K}\int_K p^{\rm D}(x)\,{\rm d}x & \text{for } K \in \mathscr{T},\\ p_\sigma^{\rm D} = \frac{1}{m_\sigma}\int_\sigma p^{\rm D}(x)\,{\rm d}m^{d-1}(x) & \text{for } \sigma \in \mathscr{E}_{\rm ext}^{\rm D}, \end{cases} \tag{3.4.12}$$

whereas the initial condition is discretized into

$$s_K^0 = \frac{1}{m_K}\int_K s^0(x)\,{\rm d}x, \qquad \text{for } K \in \mathscr{T}. \tag{3.4.13}$$

The Dirichlet boundary condition is encoded in the fluxes (3.4.11a) by setting

$$p_\sigma^n = p_\sigma^{\rm D}, \qquad \forall\sigma \in \mathscr{E}_{\rm ext}^{\rm D},\ n \geqslant 1. \tag{3.4.14}$$

Bearing in mind the definition (3.4.6) of the mirror values for $\sigma \in \mathscr{E}_{\rm ext}^{\rm N}$, the no-flux boundary condition across $\sigma \in \mathscr{E}_{\rm ext}^{\rm N}$ is automatically encoded, i.e., $F_{K\sigma}^n = 0$ for all $\sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}$, $K \in \mathscr{T}$ and $n \geqslant 1$.

In what follows, we denote by $\boldsymbol{p}^n = (p_K^n)_{K\in\mathscr{T}}$ for $1 \leqslant n \leqslant N$, and by $\boldsymbol{s}^n = (s_K^n)_{K\in\mathscr{T}}$ for $0 \leqslant n \leqslant N$. Besides, we set $\boldsymbol{p}^{\rm D} = ((p_K^{\rm D})_{K\in\mathscr{T}}, (p_\sigma^{\rm D})_{\sigma\in\mathscr{E}^{\rm D}})$.

3.4.3 Main results and organization of this chapter

The theoretical part of this chapter includes two main results. The first one, which emerges from the analysis at fixed grid, states that the schemes admits a unique solution $(\boldsymbol{p}^n)_{1\leqslant n\leqslant N}$.

Theorem 3.4.2. *For all $n \in \{1, \dots, N\}$, there exists a unique solution $\boldsymbol{p}^n$ to the scheme* (3.4.9)–(3.4.11c).

With Theorem 3.4.2 at hand, we define the approximate pressure $p_{\mathscr{T},\Delta t}$ by

$$p_{\mathscr{T},\boldsymbol{\Delta t}}(t,x) = p_K^n \qquad \text{for } (t,x) \in (t^{n-1},t^n] \times K. \tag{3.4.15a}$$

We also define the approximate saturation as

$$s_{\mathscr{T},\boldsymbol{\Delta t}} = \mathcal{S}(p_{\mathscr{T},\boldsymbol{\Delta t}}, x). \tag{3.4.15b}$$

The second main result guarantees the convergence towards a weak solution of the sequence of approximate solutions as the mesh size and the time-steps tend to 0. Let $(\mathscr{T}_m, \mathscr{E}_m, (x_K)_{K\in\mathscr{T}_m})_{m\geqslant 1}$ be a sequence of admissible discretizations of the domain Ω in the sense of Definition 3.4.1 such that

$$h_{\mathscr{T}_m} \underset{m\to\infty}{\longrightarrow} 0, \qquad \sup_{m\geqslant 1} \zeta_{\mathscr{T}_m} =: \zeta < +\infty, \tag{3.4.16}$$

where the size $h_{\mathscr{T}_m}$ and the regularity $\zeta_{\mathscr{T}_m}$ are defined in (3.4.8). Let $(\boldsymbol{\Delta t}_m)_{m\geqslant 1}$ be time discretizations of $(0,T)$ such that

$$\lim_{m\to\infty} \max_{1\leqslant n\leqslant N_m} \Delta t_m^n = 0. \tag{3.4.17}$$

Theorem 3.4.3. *There exists a weak solution $p : Q_T \to \mathbb{R}$ in the sense of Definition 3.2.1 such that, up to a subsequence,*

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \mathcal{S}(p,x) \qquad \text{a.e. in } Q_T, \tag{3.4.18a}$$

$$\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to\infty}{\longrightarrow} \Upsilon(p) \qquad \text{weakly in } L^2(Q_T). \tag{3.4.18b}$$

The rest of this chapter is outlined as follows. Section §3.5 is devoted to the numerical analysis at fixed grid. This encompasses the existence and uniqueness result stated in Theorem 3.4.2 as well as a priori estimates that will help proving Theorem 3.4.3. The convergence of the scheme, which is taken up in §3.6, relies on compactness arguments, which require a priori estimates that are uniform w.r.t. the grid. These estimates are mainly adaptations to the discrete setting of their continuous counterparts that arised in the stability analysis sketched out in §3.2. These estimates are shown in §3.6.1 to provide some compactness on the sequence of approximate solutions. In §3.6.2, we show that these compactness properties together with the a priori estimates are sufficient to identify any limit of an approximate solution as a weak solution to the problem.

Remark 3.4.4. *Theorem 3.4.3 only states the convergence of the scheme up to a subsequence. In the case where the weak solution is unique, then the whole sequence of approximate solutions would converge towards this solution. As far as we know, uniqueness of the weak solutions to Richards' equation is in general an open problem for heterogeneous media where $x \mapsto \mathcal{S}(p,x)$ is discontinuous. Uniqueness results are however available in the one-dimensional setting for a slightly more restrictive notion of solutions, cf. [41], or under additional assumptions on the nonlinearities $\eta_i, \mathcal{S}_i$, cf. [40].*

3.5 Analysis at fixed grid

3.5.1 Some uniform a priori estimates

In this section, our aim is to derive a priori estimates on the solutions to the scheme (3.4.9)–(3.4.13). These estimates will be at the core of the existence proof of a solution to the scheme. They will also play a key role in proving the convergence of the scheme.

The main estimate on which our analysis relies is a discrete counterpart of (3.2.2). We recall that a_σ is the transmissivity introduced in (3.4.5).

Proposition 3.5.1. *There exist two constants C_1, C_2 depending only on λ, μ, p^{D}, ψ, ζ, Ω, T, ϕ, and $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$ such that*

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^n - p_{K\sigma}^n)^2 \leqslant C_1, \tag{3.5.1a}$$

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (\vartheta_K^n - \vartheta_{K\sigma}^n)^2 \leqslant C_2. \tag{3.5.1b}$$

In (3.5.1), the relationship between σ and K is to be understood as follows. For an inner edge $\sigma \in \mathscr{E}_{\mathrm{int}}$, although it can be written as $\sigma = K|L$ or $L|K$, only one of these contributes to the sum. For a boundary edge $\sigma \in \mathscr{E}_{\mathrm{ext}}$, there is only one cell K such that $\sigma \in \mathscr{E}_K$, so there is no ambiguity in the sum.

Proof. Multiplying (3.4.10) by $\Delta t^n (p_K^n - p_K^{\mathrm{D}})$, summing over $K \in \mathscr{T}$ and $n \in \{1, \dots, N\}$, and carrying out discrete integration by parts yield

$$\mathtt{A} + \mathtt{B} = 0, \tag{3.5.2}$$

where we have set

$$\mathtt{A} = \sum_{n=1}^{N} \sum_{K\in\mathscr{T}} m_K \phi_K (s_K^n - s_K^{n-1})(p_K^n - p_K^{\mathrm{D}}), \tag{3.5.3a}$$

$$\mathtt{B} = \sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (\vartheta_K^n - \vartheta_{K\sigma}^n)(p_K^n - p_K^{\mathrm{D}} - p_{K\sigma}^n + p_{K\sigma}^{\mathrm{D}}). \tag{3.5.3b}$$

The discrete energy density function $\mathfrak{E}_K : [0,1] \to \mathbb{R}_+$, defined by means of the notation (3.4.7) from the functions $f_i = \mathfrak{E}_i$ introduced in (3.1.7), is convex by construction. Consequently,

$$\mathfrak{E}_K(s_K^{n-1}) - \mathfrak{E}_K(s_K^n) \geqslant \mathfrak{E}_K'(s_K^n)(s_K^{n-1} - s_K^n) = \phi_K(p_K^n - p_K^{\mathrm{D}})(s_K^{n-1} - s_K^n).$$

Therefore, the quantity $\mathtt{A}$ of (3.5.3a) can be bounded below by

$$\begin{aligned}\mathtt{A} &\geqslant \sum_{n=1}^{N} \sum_{K\in\mathscr{T}} m_K (\mathfrak{E}_K(s_K^n) - \mathfrak{E}_K(s_K^{n-1})) \\ &= \sum_{K\in\mathscr{T}} m_K (\mathfrak{E}_K(s_K^N) - \mathfrak{E}_K(s_K^0)) \geqslant -C_{\mathtt{A}},\end{aligned} \tag{3.5.4}$$

the last inequality being a consequence of the boundedness of $\mathfrak{E}_K$ on $[0,1]$.

Writing $\vartheta = p + \psi$ and expanding each summand of (3.5.3b), we can split $\mathtt{B}$ into

$$\mathtt{B} = \mathtt{B}_1 + \mathtt{B}_2 + \mathtt{B}_3,$$

with

$$\begin{aligned}
\mathtt{B}_1 &= \sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^n - p_{K\sigma}^n)^2,\\
\mathtt{B}_2 &= \sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^n - p_{K\sigma}^n)(\psi_K - \psi_{K\sigma} - p_K^{\mathrm{D}} + p_{K\sigma}^{\mathrm{D}}),\\
\mathtt{B}_3 &= -\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (\psi_K - \psi_{K\sigma})(p_K^{\mathrm{D}} - p_{K\sigma}^{\mathrm{D}}).
\end{aligned}$$

It follows from [73, Lemma 9.4] and from the boundedness of η that there exists a constant C depending only on λ, μ, $\zeta_{\mathscr{T}}$ and Ω such that

$$\sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^{\mathrm{D}} - p_{K\sigma}^{\mathrm{D}})^2 \leqslant C \left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d}^2, \tag{3.5.5a}$$

$$\sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (\psi_K - \psi_{K\sigma})^2 \leqslant C \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}^2. \tag{3.5.5b}$$

Thanks to these estimates and to the Cauchy-Schwarz inequality, we have

$$\mathtt{B}_3 \geqslant -CT \left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d} \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}.$$

On the other hand, Young's inequality provides

$$\mathtt{B}_2 \geqslant -\frac{1}{2}\mathtt{B}_1 - CT\left(\left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d}^2 + \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}^2\right).$$

Hence,

$$\mathtt{B} \geqslant \frac{1}{2}\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^n - p_{K\sigma}^n)^2 - C_{\mathtt{B}}, \tag{3.5.6}$$

by setting $C_{\mathtt{B}} = CT\left(\left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d}^2 + \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}^2 + \left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d} \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}\right)$. Inserting (3.5.4) and (3.5.6) into (3.5.2), we recover (3.5.1a) with $C_1 = 2(C_{\mathtt{A}} + C_{\mathtt{B}})$. From (3.5.1a), we can deduce (3.5.1b) by elementary manipulations. □

So far, we have not used the upwind choice (3.4.11c) for the mobilities η_σ^n. This will be done in the next lemma, where we derive a more useful variant of estimate (3.5.1a), in which η_σ^n is replaced by $\overline{\eta}_\sigma^n$ defined below. In a homogeneous medium, $\overline{\eta}_\sigma^n \geqslant \eta_\sigma^n$ so that the new estimate (3.5.8) seems to be stronger than (3.5.1a).

We begin by introducing the functions $\breve{\eta}_\sigma : \mathbb{R} \to (0, 1/\mu]$ defined for $\sigma \in \mathscr{E}$ by

$$\breve{\eta}_\sigma(p) = \min\left\{\eta_K \circ \mathcal{S}_K(p), \eta_{K\sigma} \circ \mathcal{S}_{K\sigma}(p)\right\}, \qquad \forall p \in \mathbb{R}. \tag{3.5.7a}$$

By virtue of assumptions (3.1.6), each argument of the minimum function is nondecreasing and positive function of $p \in \mathbb{R}$. As a result, $\breve{\eta}_\sigma$ is also a nondecreasing and positive function of $p \in \mathbb{R}$.

Note that $\breve{\eta}_\sigma = \eta_i \circ \mathcal{S}_i$ for all $\sigma \in \mathscr{E}_i$, while for interface edges $\sigma \subset \Gamma_{i,j}$, the mere inequality $\breve{\eta}_\sigma \leqslant \eta_i \circ \mathcal{S}_i$ holds. Next, we consider the intervals

$$\mathfrak{J}_\sigma^n = [p_K^n \bot p_{K\sigma}^n,\, p_K^n \top p_{K\sigma}^n], \qquad \text{for } \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ 1 \leqslant n \leqslant N, \tag{3.5.7b}$$

with the notations $a \bot b = \min(a,b)$ and $a \top b = \max(a,b)$. At last, we set

$$\overline{\eta}_\sigma^n = \max_{p \in \mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p), \qquad \text{for } \sigma \in \mathscr{E},\ 1 \leqslant n \leqslant N. \tag{3.5.7c}$$

Lemma 3.5.2. *There exists a constant C_3 depending on the same data as C_1 such that*

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \overline{\eta}_\sigma^n \,(p_K^n - p_{K\sigma}^n)^2 \leqslant C_3. \tag{3.5.8}$$

Proof. We partition the set $\mathscr{E}$ of edges into three subsets, namely,

$$\mathscr{E}_+^n = \big\{\sigma \mid \vartheta_K^n > \vartheta_{K\sigma}^n\big\}, \quad \mathscr{E}_-^n = \big\{\sigma \mid \vartheta_K^n < \vartheta_{K\sigma}^n\big\}, \quad \mathscr{E}_0^n = \big\{\sigma \mid \vartheta_K^n = \vartheta_{K\sigma}^n\big\}.$$

Invoking $\breve{\eta}_\sigma = \min(\eta_K \circ \mathcal{S}_K,\, \eta_{K\sigma} \circ \mathcal{S}_{K\sigma})$, we can minorize the left-hand side of (3.5.1a) to obtain

$$\sum_{n=1}^{N} \Delta t^n \Big[\sum_{\sigma \in \mathscr{E}_+^n} a_\sigma \lambda_\sigma \breve{\eta}_\sigma(p_K^n)(p_K^n - p_{K\sigma}^n)^2 + \sum_{\sigma \in \mathscr{E}_-^n} a_\sigma \lambda_\sigma \breve{\eta}_\sigma(p_{K\sigma}^n)(p_K^n - p_{K\sigma}^n)^2 + \sum_{\sigma \in \mathscr{E}_0^n} a_\sigma \lambda_\sigma \tfrac{1}{2}(\breve{\eta}_\sigma(p_K^n) + \breve{\eta}_\sigma(p_{K\sigma}^n))(p_K^n - p_{K\sigma}^n)^2 \Big] \leqslant C_1.$$

Starting from this inequality and using the boundedness of η_i and ψ, we can readily show that there exists a constant C depending on the same data as C_1 such that

$$\mathtt{D}_1 := \sum_{n=1}^{N} \Delta t^n \Big[\sum_{\sigma \in \mathscr{E}_+^n} a_\sigma \lambda_\sigma \breve{\eta}_\sigma(p_K^n)(p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) + \sum_{\sigma \in \mathscr{E}_-^n} a_\sigma \lambda_\sigma \breve{\eta}_\sigma(p_{K\sigma}^n)(p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) \Big] \leqslant C,$$

in which the sum over $\mathscr{E}_0^n$ was omitted because all of its summands vanish. Simlarly to what was pointed out in equation 2.9 in [5], we notice that since η_σ is nondecreasing w.r.t. p, it is straightforward to check that the definition

$$\breve{\eta}_\sigma^n := \begin{cases} \breve{\eta}_\sigma(p_K^n) & \text{if } \vartheta_K^n > \vartheta_{K\sigma}^n, \\ \frac{1}{2}(\breve{\eta}_\sigma(p_K^n) + \breve{\eta}_\sigma(p_{K\sigma}^n)) & \text{if } \vartheta_K^n = \vartheta_{K\sigma}^n, \\ \breve{\eta}_\sigma(p_{K\sigma}^n) & \text{if } \vartheta_K^n < \vartheta_{K\sigma}^n \end{cases} \tag{3.5.9}$$

exactly amounts to

$$\breve{\eta}_\sigma^n = \begin{cases} \max_{p \in \mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p) & \text{if } (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) > 0, \\ \frac{1}{2}(\breve{\eta}_\sigma(p_K^n) + \breve{\eta}_\sigma(p_{K\sigma}^n)) & \text{if } (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) = 0, \\ \min_{p \in \mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p) & \text{if } (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) < 0. \end{cases} \tag{3.5.10}$$

Taking advantage of this equivalence, we can transform $\mathtt{D}_1$ into

$$\mathtt{D}_1 = \sum_{n=1}^{N} \Delta t^n \Big[\sum_{\sigma\in\mathscr{E}_>^n} a_\sigma\lambda_\sigma \max_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) + \sum_{\sigma\in\mathscr{E}_<^n} a_\sigma\lambda_\sigma \min_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) \Big] \leqslant C, \quad (3.5.11)$$

where $\mathscr{E}_>^n = \{\sigma \,|\, (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) > 0\}$ and $\mathscr{E}_<^n = \{\sigma \,|\, (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) < 0\}$. The second sum over $\mathscr{E}_<^n$ contains only negative summands and can be further minorized if $\min_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma$ is replaced by $\max_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma$. In other words,

$$\mathtt{D}_2 := \sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma\lambda_\sigma \overline{\eta}_\sigma^n (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) \leqslant \mathtt{D}_1 \leqslant C.$$

Writing $\vartheta = p + \psi$, expanding each summand of $\mathtt{D}_2$ and applying Young's inequality, we end up with

$$\frac{1}{2} \sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma\lambda_\sigma \overline{\eta}_\sigma^n \left[(p_K^n - p_{K\sigma}^n)^2 - (\psi_K^n - \psi_{K\sigma}^n)^2 \right] \leqslant \mathtt{D}_2 \leqslant C.$$

Estimate (3.5.8) finally follows from the boundedness of η, $1/\lambda$ and ψ. □

The above lemma has several important consequences for the analysis. Let us start with discrete counterparts to estimations (3.2.4) and (3.2.5).

Corollary 3.5.3. *Let C_3 be the constant in Lemma 3.5.2. Then,*

$$\sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{\sigma\in\mathscr{E}_i} a_\sigma (\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))^2 \leqslant C_3, \quad (3.5.12a)$$

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma (\Upsilon(p_K^n) - \Upsilon(p_{K\sigma}^n))^2 \leqslant C_3. \quad (3.5.12b)$$

Moreover, there exists two constants C_4, C_5 depending on the same data as C_1 and additionnally on $\|\sqrt{\eta_i \circ \mathcal{S}_i}\|_{L^1(\mathbb{R}_-)}$, $1 \leqslant i \leqslant I$, such that

$$\sum_{n=1}^{N} \Delta t^n \sum_{K\in\mathscr{T}} m_K |\Upsilon(p_K^n)|^2 \leqslant C_4, \quad (3.5.13a)$$

$$\sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{K\in\mathscr{T}_i} m_K |\Theta_i(p_K^n)|^2 \leqslant C_5. \quad (3.5.13b)$$

Proof. Consider those edges $\sigma \in \mathscr{E}_i$ —defined in (3.4.3a)— corresponding to some fixed $i \in \{1, \dots, I\}$, for which $\breve{\eta}_\sigma = \eta_i \circ \mathcal{S}_i = |\Theta_i'|^2$ and $\overline{\eta}_\sigma^n = \max_{\mathfrak{J}_\sigma^n} |\Theta_i'|^2$ due to (3.2.3a). By summing the elementary inequality

$$(\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))^2 \leqslant \overline{\eta}_\sigma^n \, (p_K^n - p_{K\sigma}^n)^2,$$

over $\sigma \in \mathscr{E}_i$, $i \in \{1, \dots, I\}$ and $n \in \{1, \dots, N\}$ using appropriate weights, we get

$$\sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{\sigma \in \mathscr{E}_i} a_\sigma (\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))^2 \leqslant \sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{\sigma \in \mathscr{E}_i} a_\sigma \overline{\eta}_\sigma^n \, (p_K^n - p_{K\sigma}^n)^2,$$

whose right-hand side is obviously less than C_3, thanks to (3.5.8). This proves (3.5.12a).

Similarly, the respective definitions of $\overline{\eta}_\sigma^n$ and Υ have been tailored so that $\max_{\mathfrak{I}_\sigma^n} |\Upsilon'|^2 \leqslant \overline{\eta}_\sigma^n$ for all $\sigma \in \mathscr{E}$. As a consequence,

$$(\Upsilon(p_K^n) - \Upsilon(p_{K\sigma}^n))^2 \leqslant \overline{\eta}_\sigma^n (p_K^n - p_{K\sigma}^n)^2.$$

Summing these inequalities over $\sigma \in \mathscr{E}$ and $n \in \{1, \dots, N\}$ with appropriate weights and invoking (3.5.8), we prove (3.5.12b).

The argument for (3.5.13a) is subtler. Starting from the basic inequality

$$\begin{aligned}(\Upsilon(p_K^n) - \Upsilon(p_K^{\mathrm{D}}) - \Upsilon(p_{K\sigma}^n) + \Upsilon(p_{K\sigma}^{\mathrm{D}}))^2 \\ \leqslant 2(\Upsilon(p_K^n) - \Upsilon(p_{K\sigma}^n))^2 + 2(\Upsilon(p_K^{\mathrm{D}}) - \Upsilon(p_{K\sigma}^{\mathrm{D}}))^2,\end{aligned}$$

we apply the discrete Poincaré inequality of [73, Lemma 9.1] —which is legitimate since Γ^{D} has positive measure— followed by [73, Lemma 9.4] to obtain

$$\sum_{n=1}^{N} \Delta t^n \sum_{K \in \mathscr{T}} m_K (\Upsilon(p_K^n) - \Upsilon(p_K^{\mathrm{D}}))^2 \leqslant 2C_{\mathrm{P},\mathscr{T}} \big(C_3 + C_\zeta T \|\Upsilon'\|_\infty \|\nabla p^{\mathrm{D}}\|^2\big),$$

where $C_{\mathrm{P},\mathscr{T}}$ denotes the discrete Poincaré constant, and C_ζ is the quantity appearing in [73, Lemma 9.4] and only depends on $\zeta_{\mathscr{T}}$. This entails (3.5.13a) with $C_4 = 4C_{\mathrm{P},\mathscr{T}} \big(C_3 + C_\zeta T \|\Upsilon'\|_\infty \|\nabla p^{\mathrm{D}}\|^2\big) + 2m_\Omega T \|\Upsilon(p^{\mathrm{D}})\|_\infty^2$.

The last estimate (3.5.13b) results from the comparison (3.2.6) of the nonlinearities Θ_i and Υ. □

The purpose of the next lemma is to work out a weak estimate on the discrete counterpart of $\partial_t s$, which will lead to compactness properties in §3.6.1. For $\varphi \in C_c^\infty(Q_T)$, let

$$\varphi_K^n = \frac{1}{m_K} \int_K \varphi(t^n, x) \, \mathrm{d}x, \qquad \forall K \in \mathscr{T}, \ 1 \leqslant n \leqslant N.$$

Lemma 3.5.4. *There exists a constant C_6 depending on the same data as C_1 such that*

$$\sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \phi_K (s_K^n - s_K^{n-1}) \varphi_K^n \leqslant C_6 \|\nabla \varphi\|_{L^\infty(Q_T)^d}, \quad \forall \varphi \in C_c^\infty(Q_T). \tag{3.5.14}$$

Proof. Multiplying (3.4.10) by $\Delta t^n \, \varphi_K^n$, summing over $K \in \mathscr{T}$ and $n \in \{1, \cdots, N\}$ and carrying out discrete integration by parts, we end up with

$$\mathtt{A} := \sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \phi_K (s_K^n - s_K^{n-1}) \varphi_K^n = -\sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n (\vartheta_K^n - \vartheta_{K\sigma}^n)(\varphi_K^n - \varphi_{K\sigma}^n).$$

Applying the Cauchy-Schwarz inequality and using (3.5.1b), we get

$$\mathtt{A}^2 \leqslant C_2 \frac{\max_i \lambda_i}{\mu} \sum_n \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma (\varphi_K^n - \varphi_{K\sigma}^n)^2. \tag{3.5.15}$$

The conclusion (3.5.14) is then reached by means of the property (see [12, Section 4.4])

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma(\varphi_K^n - \varphi_{K\sigma}^n)^2 \leqslant C\|\nabla\varphi\|^2_{L^\infty(Q_T)^d}$$

for some C depending only on Ω, T and the mesh regularity $\zeta_{\mathscr{T}}$. □

3.5.2 Existence of a solution to the scheme

The statements of the previous section are all uniform w.r.t. the mesh and are meant to help us passing to the limit in the next section. In contrast, the next lemma provides a bound on the pressure that depends on the mesh size and on the time-step. This property is needed in the process of ensuring the existence of a solution to the numerical scheme.

Lemma 3.5.5. *There exist two constants C_7, C_8 depending on $\mathscr{T}$, Δt^n as well as on the data of the continuous model λ, μ, p^{D}, ψ, ζ, Ω, T, ϕ, $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$ and $\|\sqrt{\eta_i\circ\mathcal{S}_i}\|_{L^1(\mathbb{R}_-)}$, $1\leqslant i\leqslant I$, such that*

$$-C_7 \leqslant p_K^n \leqslant C_8, \quad \forall K\in\mathscr{T},\ n\in\{1,\dots,N\}. \tag{3.5.16}$$

Proof. From (3.5.13a) and from $\Upsilon(p) = p\sqrt{\min_i \lambda_i/\mu}$ for $p\geqslant 0$, we deduce that

$$p_K^n \leqslant \sqrt{\frac{\mu C_4}{\Delta t^n m_K \min_i \lambda_i}}, \qquad \forall K\in\mathscr{T},\ 1\leqslant n\leqslant N.$$

Hence, the upper-bound C_8 is found by maximizing the right-hand side over $K\in\mathscr{T}$ and $n\in\{1,\dots,N\}$.

To show that p_K^n is bounded from below, we employ a strategy that was developed in [43] and extended to the case of Richards' equation in [5, Lemma 3.10]. From (3.4.12), (3.4.14) and the boundedness of p^{D}, it is easy to see that

$$p_\sigma^n \geqslant \inf_{x\in\partial\Omega} p^{\mathrm{D}}(x), \quad \forall\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}.$$

Estimate (3.5.8) then shows that for all $K\in\mathscr{T}$ such that $\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}\neq\varnothing$, we have

$$p_K^n \geqslant p_\sigma^n - \sqrt{\frac{C_3}{\Delta t^n a_\sigma \breve{\eta}_\sigma(p_\sigma^n)}} =: \pi_K^n, \quad \forall\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}.$$

The quantity π_K^n is well-defined, since $\breve{\eta}_\sigma(p_\sigma^n)>0$ for $p_\sigma^n>-\infty$, and does not depend on time, as p^{D} does not either. Furthermore, if p_K^n is bounded from below by some π_K, then the pressure in all its neighboring cells $L\in\mathscr{T}$ such that $\sigma = K|L\in\mathscr{E}_K$ is bounded from below by

$$p_L^n \geqslant \pi_K^n - \sqrt{\frac{C_3}{\Delta t^n a_\sigma \breve{\eta}_\sigma(\pi_K^n)}} =: \pi_L^n.$$

Again, π_L^n is well-defined owing to $\breve{\eta}_\sigma(\pi_K^n)>0$. Since the mesh is finite and since the domain is connected, only a finite number of edge-crossings is required to create a path from a Dirichlet boundary edge $\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}$ to any prescribed cell $K\in\mathscr{T}$. Hence, the lower bound C_7 is found by minimizing π_K^n over $K\in\mathscr{T}$ and $n\in\{1,\dots,N\}$. □

Lemma 3.5.5 is a crucial step in the proof of the existence of a solution $\boldsymbol{p}^n = (p_K^n)_{K\in\mathscr{T}}$ to the scheme (3.4.9)–(3.4.14).

Proposition 3.5.6. *Given $\boldsymbol{s}^{n-1} = (s_K^{n-1})_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$, there exists a solution $\boldsymbol{p}^n \in \mathbb{R}^{\mathscr{T}}$ to the scheme* (3.4.9)–(3.4.14)*.*

The proof relies on a standard topological degree argument and is omitted here. However, we make the homotopy explicit for readers' convenience. Let $\gamma \in [0,1]$ be the homotopy parameter. We define the nondecreasing functions $\eta_i^{(\gamma)} : [0,1] \to \mathbb{R}_+$ by setting $\eta_i^{(\gamma)}(s) = (1-\gamma)/\mu + \gamma\eta_i(s)$ for $s \in [0,1]$, and we seek a solution $\boldsymbol{p}^{(\gamma)} = (p_K^{(\gamma)})_{K\in\mathscr{T}}$ to the problem

$$\gamma m_K \phi_K \frac{\mathcal{S}_K(p_K^{(\gamma)}) - s_K^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathscr{E}_K} F_{K\sigma}^{(\gamma)} = 0, \qquad K \in \mathscr{T},\ \gamma \in [0,1], \tag{3.5.17a}$$

where the fluxes $F_{K\sigma}^{(\gamma)}$ are defined by

$$F_{K\sigma}^{(\gamma)} = \frac{m_\sigma}{d_\sigma}\lambda_\sigma \eta_\sigma^{(\gamma)}\big(\vartheta_K^{(\gamma)} - \vartheta_{K\sigma}^{(\gamma)}\big), \qquad \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ \gamma \in [0,1] \tag{3.5.17b}$$

with $\vartheta^{(\gamma)} = p^{(\gamma)} + \psi$ and using the upwind mobilities

$$\eta_\sigma^{(\gamma)} = \begin{cases} \eta_K^{(\gamma)}(\mathcal{S}_K(p_K^{(\gamma)})) & \text{if } \vartheta_K^{(\gamma)} > \vartheta_{K\sigma}^{(\gamma)}, \\ \frac{1}{2}(\eta_K^{(\gamma)}(\mathcal{S}_K(p_K^{(\gamma)})) + \eta_{K\sigma}^{(\gamma)}(\mathcal{S}_{K\sigma}(p_K^{(\gamma)}))) & \text{if } \vartheta_K^{(\gamma)} = \vartheta_{K\sigma}^{(\gamma)}, \\ \eta_{K\sigma}^{(\gamma)}(\mathcal{S}_{K\sigma}(p_K^{(\gamma)})) & \text{if } \vartheta_K^{(\gamma)} < \vartheta_{K\sigma}^{(\gamma)}. \end{cases} \tag{3.5.17c}$$

At the Dirichlet boundary edges, we still set $p_\sigma^{(\gamma)} = p_\sigma^{\mathrm{D}}$. For $\gamma = 0$, the system is linear and invertible, while for $\gamma = 1$, system (3.5.17) coincides with the original system (3.4.9)–(3.4.14). A priori estimates on $\boldsymbol{p}^{(\gamma)}$ that are uniform w.r.t. $\gamma \in [0,1]$ (but not uniform w.r.t. $\mathscr{T}$ nor Δt^n) can be derived on the basis of what was exposed previously, so that one can unfold Leray-Schauder's machinery [56, 105] to prove the existence of (at least) one solution to the scheme.

3.5.3 Uniqueness of the discrete solution

To complete the proof of Theorem 3.4.2, it remains to show that the solution to the scheme is unique. This is the purpose of the following proposition.

Proposition 3.5.7. *Given $\boldsymbol{s}^{n-1} = (s_K^{n-1})_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$, the solution $\boldsymbol{p}^n \in \mathbb{R}^{\mathscr{T}}$ to the scheme* (3.4.9)–(3.4.14) *is unique.*

Proof. The proof heavily rests upon the monotonicity properties inherited from the upwind choice (3.4.11c) for the mobilities. Indeed, due to the upwind choice of the mobility, the flux $F_{K\sigma}^n$ is a function of p_K^n and $p_{K\sigma}^n$ that is nondecreasing w.r.t. p_K^n and nonincreasing w.r.t. $p_{K\sigma}^n$. Moreover, by virtue of the monotonicity of $\mathcal{S}_K$, the discrete volume balance (3.4.10) can be cast under the abstract form

$$\mathcal{H}_K^n(p_K^n, (p_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) = 0, \qquad \forall K \in \mathscr{T}, \tag{3.5.18}$$

where $\mathcal{H}_K^n$ is nondecreasing w.r.t its first argument p_K^n and nonincreasing w.r.t each of the remaining variables $(p_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}$.

Let $\widetilde{\boldsymbol{p}}^n = (\widetilde{p}_K^n)_{K\in\mathscr{T}}$ be another solution to the system (3.4.9)–(3.4.14), i.e.,

$$\mathcal{H}_K^n(\widetilde{p}_K^n, (\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) = 0, \qquad \forall K\in\mathscr{T}. \tag{3.5.19}$$

The nonincreasing behavior of $\mathcal{H}_K^n$ w.r.t. all its variables except the first one implies that

$$\mathcal{H}_K^n(p_K^n, (p_{K\sigma}^n\top\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) \leqslant 0, \qquad \mathcal{H}_K^n(\widetilde{p}_K^n, (p_{K\sigma}^n\top\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) \leqslant 0,$$

for all $K\in\mathscr{T}$, where $a\top b = \max(a,b)$. Since $p_K^n\top\widetilde{p}_K^n$ is either equal to p_K^n or to $\widetilde{p}_K^n$, we infer from the above inequalities that

$$\mathcal{H}_K^n(p_K^n\top\widetilde{p}_K^n, (p_{K\sigma}^n\top\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) \leqslant 0, \qquad \forall K\in\mathscr{T}. \tag{3.5.20}$$

By a similar argument, we can show that

$$\mathcal{H}_K^n(p_K^n\bot\widetilde{p}_K^n, (p_{K\sigma}^n\bot\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}) \geqslant 0, \qquad \forall K\in\mathscr{T}, \tag{3.5.21}$$

where $a\bot b = \min(a,b)$. Subtracting (3.5.21) from (3.5.20) and summing over $K\in\mathscr{T}$, we find

$$\sum_{K\in\mathscr{T}} m_K\phi_K\frac{|s_K^n - \widetilde{s}_K^n|}{\Delta t^n} + \sum_{\sigma\in\mathscr{E}_{\rm ext}^{\rm D}} a_\sigma\lambda_\sigma \mathtt{R}_\sigma^n \leqslant 0, \tag{3.5.22}$$

where $s_K^n = \mathcal{S}_K(p_K^n)$, $\widetilde{s}_K^n = \mathcal{S}_K(\widetilde{p}_K^n)$ and

$$\begin{aligned}\mathtt{R}_\sigma^n = {}& \eta_K(s_K^n\top\widetilde{s}_K^n)(\vartheta_K^n\top\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+ - \eta_K(s_\sigma^n)(\vartheta_\sigma^n - \vartheta_K^n\top\widetilde{\vartheta}_K^n)^+\\ & - \eta_K(s_K^n\bot\widetilde{s}_K^n)(\vartheta_K^n\bot\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+ + \eta_K(s_\sigma^n)(\vartheta_\sigma^n - \vartheta_K^n\bot\widetilde{\vartheta}_K^n)^+,\end{aligned} \tag{3.5.23}$$

with $s_\sigma^n = \mathcal{S}_K(p_\sigma^n)$. The top line of (3.5.23) expresses the upwinded flux of (3.5.20), while the bottom line of (3.5.23) is the opposite of the upwinded flux of (3.5.21). Note that, since $p_\sigma^n = p_\sigma^{\rm D}$ is prescribed at $\sigma\in\mathscr{E}_{\rm ext}^{\rm D}$, we have $\vartheta_\sigma^n = \vartheta_\sigma^n\top\widetilde{\vartheta}_\sigma^n = \vartheta_\sigma^n\bot\widetilde{\vartheta}_\sigma^n$. Upon inspection of the rearrangement

$$\begin{aligned}\mathtt{R}_\sigma^n = {}& [\eta_K(s_K^n\top\widetilde{s}_K^n) - \eta_K(s_K^n\bot\widetilde{s}_K^n)](\vartheta_K^n\top\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+\\ & + \eta_K(s_K^n\bot\widetilde{s}_K^n)[(\vartheta_K^n\top\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+ - (\vartheta_K^n\bot\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+]\\ & + \eta_K(s_\sigma^n)\big[(\vartheta_\sigma^n - \vartheta_K^n\bot\widetilde{\vartheta}_K^n)^+ - (\vartheta_\sigma^n - \vartheta_K^n\top\widetilde{\vartheta}_K^n)^+\big],\end{aligned} \tag{3.5.24}$$

it is trivial that $\mathtt{R}_\sigma^n \geqslant 0$. As a consequence, (3.5.22) implies that $\mathtt{R}_\sigma^n = 0$ for all $\sigma\in\mathscr{E}_{\rm ext}^{\rm D}$ and that $s_K^n = \widetilde{s}_K^n$ for all $K\in\mathscr{T}$. At this stage, however, we cannot yet claim that $p_K^n = \widetilde{p}_K^n$, as the function $\mathcal{S}_K$ is not invertible.

Taking into account $s_K^n = \widetilde{s}_K^n$, the residue (3.5.24) becomes

$$\begin{aligned}\mathtt{R}_\sigma^n = {}& \eta_K(s_K^n)[(\vartheta_K^n\top\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+ - (\vartheta_K^n\bot\widetilde{\vartheta}_K^n - \vartheta_\sigma^n)^+]\\ & + \eta_K(s_\sigma^n)\big[(\vartheta_\sigma^n - \vartheta_K^n\bot\widetilde{\vartheta}_K^n)^+ - (\vartheta_\sigma^n - \vartheta_K^n\top\widetilde{\vartheta}_K^n)^+\big],\end{aligned} \tag{3.5.25}$$

which can be lower-bounded by

$$\mathtt{R}_\sigma^n \geqslant \min(\eta_K(s_K^n),\, \eta_K(s_\sigma^n))|\vartheta_K^n - \widetilde{\vartheta}_K^n| \tag{3.5.26}$$

thanks to the algebraic identities $a^+ - (-a)^+ = a$ and $a\top b - a\bot b = |a-b|$. In view of the lower-bound on the discrete pressures of Lemma 3.5.5, we deduce from (3.1.6b) that $s_K^n > 0$ and

$\widetilde{s}_K^n > 0$. The increasing behavior of η_K implies, in turn, that $\eta_K(s_K^n) > 0$ and $\eta_K(\widetilde{s}_K^n) > 0$. Therefore, the conjunction of $\mathtt{R}_\sigma^n = 0$ and (3.5.26) yields $\vartheta_K^n = \widetilde{\vartheta}_K^n$ and hence $p_K^n = \widetilde{p}_K^n$ for all cells K having a Dirichlet boundary edge, i.e., $\mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm D} \neq \varnothing$.

It remains to check that $p_K^n = \widetilde{p}_K^n$, or equivalently $\vartheta_K^n = \widetilde{\vartheta}_K^n$ for those cells $K \in \mathscr{T}$ that are far away from the Dirichlet part of the boundary. Subtracting (3.5.19) from (3.5.18) and recalling that $s_K^n = \widetilde{s}_K^n$, we arrive at

$$\sum_{\sigma\in\mathscr{E}_K} a_\sigma\lambda_\sigma\Big\{\eta_K(s_K^n)\big[(\vartheta_K^n - \vartheta_{K\sigma}^n)^+ - (\widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n)^+\big] + \eta_{K\sigma}(s_{K\sigma}^n)\big[(\widetilde{\vartheta}_{K\sigma}^n - \widetilde{\vartheta}_K^n)^+ - (\vartheta_{K\sigma}^n - \vartheta_K^n)^+\big]\Big\} = 0. \tag{3.5.27}$$

Consider a cell $K \in \mathscr{T}$ where $\vartheta_K^n - \widetilde{\vartheta}_K^n$ achieves its maximal value, i.e.,

$$\vartheta_K^n - \widetilde{\vartheta}_K^n \geqslant \vartheta_L^n - \widetilde{\vartheta}_L^n, \qquad \forall L \in \mathscr{T}. \tag{3.5.28}$$

This entails that

$$\vartheta_K^n - \vartheta_{K\sigma}^n \geqslant \widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n, \qquad \forall \sigma \in \mathscr{E}_K,$$

so that the two brackets in the right-hand side of (3.5.27) are nonnegative. In fact, they both vanish by the positivity of $\eta_K(s_K^n)$ and $\eta_{K\sigma}(s_{K\sigma}^n)$. As a result, $\vartheta_K^n - \vartheta_{K\sigma}^n = \widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n$ for all $\sigma \in \mathscr{E}_K$. This implies that $\vartheta_K^n - \widetilde{\vartheta}_K^n = \vartheta_L^n - \widetilde{\vartheta}_L^n$ for all the cells $L \in \mathscr{T}$ sharing an edge $\sigma = K|L$ with K, and thus that the cell L also achieves the maximality condition (3.5.28). The process can then be repeated over and over again. Since Ω is connected, we deduce that $\vartheta_K^n - \widetilde{\vartheta}_K^n$ is constant over $K \in \mathscr{T}$. The constant is finally equal to zero since $\vartheta_K^n = \widetilde{\vartheta}_K^n$ on the cells having a Dirichlet edge. □

3.6 Convergence analysis

Once existence and uniqueness of the discrete solution have been settled, the next question to be addressed is the convergence of the discrete solution towards a weak solution of the continuous problem, as the mesh-size and the time-step are progressively refined. In accordance with the general philosophy expounded in [73], the proof is built on compactness arguments. We start by highlighting compactness properties in §3.6.1, before identifying the limit values as weak solutions in §3.6.2.

3.6.1 Compactness properties

Let us define $G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} : Q_T \to \mathbb{R}^d$ and $J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} : Q_T \to \mathbb{R}^d$ by

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} d\dfrac{\Theta_i(p_{K\sigma}^n) - \Theta_i(p_K^n)}{d_\sigma} n_{K\sigma}, & \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6.1}$$

for $\sigma \in \mathscr{E}_{i,m}$, $1 \leqslant n \leqslant N_m$ and, respectively,

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = d\frac{\Upsilon(p_{K\sigma}^n) - \Upsilon(p_K^n)}{d_\sigma} n_{K\sigma}, \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \tag{3.6.2}$$

for $\sigma \in \mathscr{E}_m$, $1 \leqslant n \leqslant N_m$. We remind that $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \mathcal{S}(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, x)$ is the sequence of approximate saturation fields computed from that of approximate pressure fields $p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$ by (3.4.15b).

Proposition 3.6.1. *There exists a measurable function $p : Q_T \to \mathbb{R}$ such that $\Upsilon(p) - \Upsilon(p^{\rm D}) \in L^2((0,T);V)$ and $\Theta_i(p) \in L^2((0,T);H^1(\Omega_i))$, $1 \leqslant i \leqslant I$, such that, up to a subsequence,*

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \mathcal{S}(p,x) \qquad \text{a.e. in } Q_T, \tag{3.6.3a}$$

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \nabla\Theta_i(p) \qquad \text{weakly in } L^2(Q_{i,T})^d, \tag{3.6.3b}$$

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \nabla\Upsilon(p) \qquad \text{weakly in } L^2(Q_T)^d. \tag{3.6.3c}$$

Proof. We know from Corollary 3.5.3 that $\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ and $\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ are bounded w.r.t. m in $L^2(Q_{i,T})$ and $L^2(Q_T)$ respectively, while $G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ and $J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ are respectively bounded in $L^2(Q_{i,T})^d$ and $L^2(Q_T)^d$. In particular, there exist $\widehat{\Theta}_i \in L^2(Q_{i,T})$, $\widehat{\Upsilon} \in L^2(Q_T)$, $J \in L^2(Q_{i,T})^d$, and $J \in L^2(Q_T)^d$ such that

$$\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to+\infty}{\longrightarrow} \widehat{\Theta}_i \qquad \text{weakly in } L^2(Q_{i,T}), \tag{3.6.4a}$$

$$\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to+\infty}{\longrightarrow} \widehat{\Upsilon} \qquad \text{weakly in } L^2(Q_T), \tag{3.6.4b}$$

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} G \qquad \text{weakly in } L^2(Q_{i,T})^d, \tag{3.6.4c}$$

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} J \qquad \text{weakly in } L^2(Q_T)^d. \tag{3.6.4d}$$

Establishing that $\widehat{\Theta}_i \in L^2((0,T);H^1(\Omega_i))$ and $\widehat{\Upsilon} \in L^2((0,T);H^1(\Omega))$ with $G = \nabla\widehat{\Theta}_i$ and $J = \nabla\widehat{\Upsilon}$ is now classical, see for instance [70, Lemma 2] or [50, Lemma 4.4].

The key points of this proof are the identification $\widehat{\Theta}_i = \Theta_i(p)$ and $\widehat{\Upsilon} = \Upsilon(p)$ for some measurable p, as well as the proofs of the almost everywhere convergence property (3.6.3a). The identification of the limit and the almost everywhere convergence can be handled simultaneously by using twice [12, Theorem 3.9], once for $\Theta_i(p)$ and once for $\Upsilon(p)$. More precisely, Lemma 3.5.4 provides a control on the time variations of the approximate saturation $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$, whereas Corollary 3.5.3 provides some compactness w.r.t. space on $\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ and $\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$. Using further that $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \mathcal{S}_i \circ \Theta_i^{-1}\left(\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})\right)$ with $\mathcal{S}_i \circ \Theta_i^{-1}$ nondecreasing and continuous, then one infers from [12, Theorem 3.9] that

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \mathcal{S}_i \circ \Theta_i^{-1}(\widehat{\Theta}_i) \quad \text{a.e. in } Q_{i,T}.$$

Let $p = \Theta_i^{-1}(\widehat{\Theta}_i)$. Then, (3.6.3a) and (3.6.3b) hold. Proving (3.6.3a) and (3.6.3c) is similar, and the properties (3.6.3) can be assumed to hold for the same function p up to the extraction of yet another subsequence.

Finally, by applying the arguments developed in [30, §4.2], we show that $\Upsilon(p)$ and $\Upsilon(p^{\rm D})$ share the same trace on $(0,T)\times\Gamma^{\rm D}$, hence $\Upsilon(p) - \Upsilon(p^{\rm D}) \in L^2((0,T);V)$. □

Let us now define

$$\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \eta_\sigma^n \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma \tag{3.6.5}$$

for $\sigma \in \mathscr{E}_m$, $1 \leqslant n \leqslant N_m$.

Lemma 3.6.2. *Up to a subsequence, the function p whose existence is guaranteed by Proposition 3.6.1 satisfies*

$$\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \eta(\mathcal{S}(p,x)) \qquad \text{in } L^q(Q_T),\ 1 \leqslant q < +\infty. \tag{3.6.6}$$

Proof. Because of (3.6.3a), $\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \eta(s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, x)$ converges almost everywhere to $\eta(\mathcal{S}(p,x),x)$. Since η is bounded, Lebesgue's dominated convergence theorem ensures that the convergence holds in $L^q(Q_T)$ for all $q \in [1,+\infty)$. The reconstruction $\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ of the mobility is also uniformly bounded, so we have just to show that $\|\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} - \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}\|_{L^1(Q_T)} \to 0$ as $m \to +\infty$. Letting $\Delta_{K\sigma} = K \cap \Delta_\sigma$ denote the half-diamond cell, we have

$$\begin{aligned}\|\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} - \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}\|_{L^1(Q_T)} &\leqslant \sum_{n=1}^{N_m} \Delta t_m^n \sum_{K\in\mathscr{T}_m} \sum_{\sigma\in\mathscr{E}_K} m_{\Delta_{K\sigma}} |\eta_K(s_K^n) - \eta_\sigma^n| \\ &\leqslant \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)| \leqslant \sum_{i=1}^{I} \mathtt{R}_{i,m} + \mathtt{R}_{\Gamma,m},\end{aligned}$$

where

$$\mathtt{R}_{i,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)|,$$

$$\mathtt{R}_{\Gamma,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)|.$$

Let us define

$$r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = |\eta_K^n - \eta_{K\sigma}^n| = r_\sigma^n \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma,$$

then $r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ is uniformly bounded by $\|\eta\|_\infty = 1/\mu$. Therefore,

$$\mathtt{R}_{\Gamma,m} \leqslant \frac{T}{\mu} \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} m_{\Delta_\sigma} \leqslant \frac{2T\, m^{d-1}(\Gamma)}{\mu d}\, h_{\mathscr{T}_m}$$

where $h_{\mathscr{T}_m}$ is the size of $\mathscr{T}_m$ as defined in (3.4.8). Besides, for $i \in \{1,\dots,I\}$, $\eta_i \circ \mathcal{S}_i \circ \Theta_i^{-1}$ is continuous, monotone and bounded, hence uniformly continuous. This provides the existence of a modulus of continuity $\varpi_i : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varpi_i(0) = 0$ such that

$$r_\sigma^n := |\eta \circ \mathcal{S} \circ \Theta_i^{-1}(\Theta_K^n) - \eta \circ \mathcal{S} \circ \Theta_i^{-1}(\Theta_{K\sigma}^n)| \leqslant \varpi_i(|\Theta_K^n - \Theta_{K\sigma}^n|) \tag{3.6.7}$$

for $\sigma \in \mathscr{E}_{i,m}$. Therefore, if the function

$$q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} |\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n)| & \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6.8a}$$

for $\sigma \in \mathscr{E}_{i,m}$, $1 \leqslant n \leqslant N_m$, could be proven to converge to 0 almost everywhere in $Q_{i,T}$, then it would also be the case for $r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ and $\mathtt{R}_{i,m}$ as $m \to +\infty$, thanks to Lebesgue's dominated convergence theorem. Now, it follows from (3.5.12a) and from the elementary geometric relation

$$m_{\Delta_\sigma} = \frac{a_\sigma}{d} d_\sigma^2 \leqslant 4\frac{a_\sigma}{d} h_{\mathscr{T}_m}^2,$$

that

$$\|q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m}\|_{L^2(Q_{i,T})}^2 = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} |\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n)|^2 \leqslant \frac{4C_3}{d} h_{\mathscr{T}_m}^2.$$

Therefore, $q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m} \to 0$ in $L^2(Q_{i,T})$, thus also almost everywhere up to extraction of a subsequence. This provides the desired result. □

3.6.2 Identification of the limit

So far, we have exhibited some "limit" value p for the approximate solution $p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$ in Proposition 3.6.1. Next, we show that the scheme is consistent with the continuous problem by showing that any limit value is a weak solution.

Proposition 3.6.3. *The function p whose existence is guaranteed by Proposition 3.6.1 is a weak solution of the problem* (3.1.9a)–(3.1.9c) *in the sense of Definition 3.2.1.*

Proof. Let $\varphi \in C_c^\infty(\{\Omega \cup \Gamma^{\mathrm{N}}\} \times [0,T))$ and denote by $\varphi_K^n = \varphi(t_m^n, x_K)$, for all $K \in \mathscr{T}_m$ and all $n \in \{0,\dots,N_m\}$. We multiply (3.4.10) by $\Delta t_m^n \varphi_K^{n-1}$ and sum over $n \in \{1,\dots,N_m\}$ and $K \in \mathscr{T}_m$ to obtain

$$\mathtt{A}_m + \mathtt{B}_m = 0, \qquad m \geqslant 1, \tag{3.6.9}$$

where we have set

$$\mathtt{A}_m = \sum_{n=1}^{N_m} \sum_{K\in\mathscr{T}_m} m_K \phi_K (s_K^n - s_K^{n-1}) \varphi_K^{n-1}, \tag{3.6.10a}$$

$$\mathtt{B}_m = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} a_\sigma \lambda_\sigma \eta_\sigma^n (\vartheta_K^n - \vartheta_{K\sigma}^n)(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}). \tag{3.6.10b}$$

The quantity $\mathtt{A}_m$ in (3.6.10a) can be rewritten as

$$\begin{aligned}\mathtt{A}_m &= -\sum_{n=1}^{N_m} \Delta t_m^n \sum_{K\in\mathscr{T}_m} m_K \phi_K s_K^n \frac{\varphi_K^n - \varphi_K^{n-1}}{\Delta t_m^n} - \sum_{K\in\mathscr{T}_m} m_K \phi_K s_K^0 \varphi_K^0 \\ &= -\iint_{Q_T} \phi\, s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \delta\varphi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\, s^0_{\mathscr{T}_m} \varphi^0_{\mathscr{T}_m} \,\mathrm{d}x\end{aligned}$$

where

$$\begin{aligned}\delta\varphi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}(t,x) &= \frac{\varphi_K^n - \varphi_K^{n-1}}{\Delta t_m^n}, \qquad \text{if } (t,x) \in (t_m^{n-1}, t_m^n) \times K, \\ \varphi^0_{\mathscr{T}_m} &= \varphi(0,x_K) \qquad\qquad \text{if } x \in K.\end{aligned}$$

Thanks to the regularity of φ, the function $\delta\varphi_{\mathscr{T}_m,\Delta t_m}$ converges uniformly to $\partial_t\varphi$ on $\Omega \times [0,T]$. Moreover, by virtue of (3.6.3a) and the boundedness of $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_n}$ we can state that

$$\iint_{Q_T} \phi\, s_{\mathscr{T}_m,\Delta t_m} \delta\varphi_{\mathscr{T}_m,\Delta t_m} \,\mathrm{d}x\,\mathrm{d}t \underset{m\to+\infty}{\longrightarrow} \iint_{Q_T} \phi\, \mathcal{S}(p,x) \partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t,$$

and, in view of the definition (3.4.13) of $s^0_{\mathscr{T}_m}$ and of the uniform convergence of $\varphi^0_{\mathscr{T}_m}$ towards $\varphi(0,\cdot)$,

$$\int_\Omega \phi_{\mathscr{T}_m} s^0_{\mathscr{T}_m} \varphi^0_{\mathscr{T}_m} \,\mathrm{d}x \underset{m\to+\infty}{\longrightarrow} \int_\Omega \phi\, s^0 \varphi(0,\cdot) \,\mathrm{d}x.$$

From the above, we draw that

$$\lim_{m\to+\infty} \mathtt{A}_m = -\iint_{Q_T} \phi\, \mathcal{S}(p,x) \partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\, s^0 \varphi(0,\cdot) \,\mathrm{d}x. \tag{3.6.11}$$

Let us now turn our attention to the quantity $\mathtt{B}_m$ of (3.6.10b), which can be split into $\mathtt{B}_m = \mathtt{B}_m^1 + \mathtt{B}_m^2$ using

$$\mathtt{B}_m^1 = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} a_\sigma \lambda_\sigma \eta_\sigma^n (p_K^n - p_{K\sigma}^n)(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}),$$
$$\mathtt{B}_m^2 = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} a_\sigma \lambda_\sigma \eta_\sigma^n (\psi_K - \psi_{K\sigma})(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}).$$

Consider first the convective term $\mathtt{B}_m^2$. It follows from the definition of the discrete gravitational potential

$$\psi_K = -\varrho g \cdot x_K, \quad \psi_\sigma = -\varrho g \cdot x_\sigma, \qquad K \in \mathscr{T}_m,\ \sigma \in \mathscr{E}_{\mathrm{ext},m}^{\mathrm{D}}$$

and from the orthogonality of the mesh that

$$\psi_K - \psi_{K\sigma} = d_\sigma \varrho g \cdot \nu_{K\sigma}, \qquad \forall \sigma \in \mathscr{E}_K \backslash \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}},\ K \in \mathscr{T}_m.$$

Therefore, $\mathtt{B}_m^2$ can be transformed into

$$\begin{aligned}\mathtt{B}_m^2 &= \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} m_{\Delta_\sigma} \lambda_\sigma \eta_\sigma^n d \frac{\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}}{d_\sigma} n_{K\sigma} \cdot \varrho g \\ &= -\iint_{Q_T} \lambda_{\mathscr{E}_m} \eta_{\mathscr{E}_m, \boldsymbol{\Delta t}_m} H_{\mathscr{E}_m, \boldsymbol{\Delta t}_m} \cdot \varrho g \,\mathrm{d}x\,\mathrm{d}t, \end{aligned} \tag{3.6.12}$$

where

$$\begin{aligned} \lambda_{\mathscr{E}_m}(x) &= \lambda_\sigma && \text{if } x \in \Delta_\sigma,\ \sigma \in \mathscr{E}_m, \\ H_{\mathscr{E}_m, \boldsymbol{\Delta t}_m}(t,x) &= (d/d_\sigma)(\varphi_{K\sigma}^{n-1} - \varphi_K^{n-1}) n_{K\sigma} && \text{if } (t,x) \in [t_m^{n-1}, t_m^n) \times \Delta_\sigma. \end{aligned}$$

After [50, Lemma 4.4], $H_{\mathscr{E}_m, \boldsymbol{\Delta t}_m}$ converges weakly in $L^2(Q_T)^d$ towards $\nabla\varphi$, while $\lambda_{\mathscr{E}_m}$ and $\eta_{\mathscr{E}_m, \boldsymbol{\Delta t}_m}$ converge strongly in $L^4(\Omega)$ and $L^4(Q_T)$ towards λ and $\eta(\mathcal{S}(p,x))$ respectively (cf. Lemma 3.6.2). Thus, we can pass to the limit in (3.6.12) and

$$\lim_{m\to+\infty} \mathtt{B}_m^2 = -\iint_{Q_T} \lambda \eta(\mathcal{S}(p,x)) \varrho g \cdot \nabla\varphi \,\mathrm{d}x\,\mathrm{d}t. \tag{3.6.13}$$

The capillary diffusion term $\mathtt{B}_m^1$ appears to be the most difficult one to deal with. Taking inspiration from [43], we introduce the auxiliary quantity

$$\begin{aligned}\widetilde{\mathtt{B}}_m^1 &= \sum_{i=1}^{I} \widetilde{\mathtt{B}}_{i,m}^1 \\ &= \sum_{i=1}^{I} \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma \sqrt{\lambda_i \eta_\sigma^n} (\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}).\end{aligned}$$

Analogously to [61], we can define a piecewise-constant vector field $\overline{H}_{\mathscr{E}_m, \boldsymbol{\Delta t}_m}$ such that

$$\overline{H}_{\mathscr{E}_m, \boldsymbol{\Delta t}_m}(t,x) \cdot \nu_{K\sigma} = \varphi_{K\sigma}^{n-1} - \varphi_K^{n-1}, \qquad \text{if } (t,x) \in [t_m^{n-1}, t_m^n) \times \Delta_\sigma,\ \sigma \in \mathscr{E}_m,$$

and such that $\overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ converges uniformly towards $\nabla\varphi$ on $\overline{Q}_T$. Under these circumstances, $\widetilde{\mathtt{B}}^1_{i,m}$ reads

$$\widetilde{\mathtt{B}}^1_{i,m} = \int_0^T \int_{\Omega_{i,m}} \sqrt{\lambda_i \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}}\, G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \cdot \overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \,\mathrm{d}x\,\mathrm{d}t$$

where $\Omega_{i,m} = \bigcup_{\sigma\in\mathscr{E}_{i,m}} \Delta_\sigma \subset \Omega_i$. The strong convergence of $\sqrt{\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}}$ in $L^2(Q_{i,T})$ towards $\sqrt{\eta_i(\mathcal{S}_i(p))}$ directly follows from the boundedness of η_i combined with (3.6.3a). Combining this with (3.6.3b) results in

$$\widetilde{\mathtt{B}}^1_{i,m} \underset{m\to+\infty}{\longrightarrow} \iint_{Q_{i,T}} \sqrt{\lambda_i \eta_i(\mathcal{S}_i(p))} \nabla\Theta_i(p)\cdot\nabla\varphi \,\mathrm{d}x\,\mathrm{d}t = \iint_{Q_{i,T}} \nabla\Phi_i(p)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t. \tag{3.6.14}$$

Therefore, to finish the proof of Proposition 3.6.3, it only remains to check that B^1_m and $\widetilde{B}^1_m$ share the same limit. To this end, we observe that, by the triangle inequality, we have

$$|\mathtt{B}^1_m - \widetilde{\mathtt{B}}^1_m| \leqslant \mathtt{R}_{\Gamma,m} + \sum_{i=1}^I \mathtt{R}_{i,m}, \tag{3.6.15}$$

where

$$\mathtt{R}_{\Gamma,m} = \sum_{n=1}^{N_m} \Delta t^n_m \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} a_\sigma \lambda_\sigma \eta^n_\sigma |p^n_K - p^n_{K\sigma}||\varphi^{n-1}_K - \varphi^{n-1}_{K\sigma}|,$$

$$\mathtt{R}_{i,m} = \sum_{n=1}^{N_m} \Delta t^n_m \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma \sqrt{\lambda_i\eta^n_\sigma} |\Theta_i(p^n_K) - \Theta_i(p^n_{K\sigma}) - \sqrt{\lambda_i\eta^n_\sigma}(p^n_K - p^n_{K\sigma})||\varphi^{n-1}_K - \varphi^{n-1}_{K\sigma}|.$$

Applying the Cauchy-Schwarz inequality and using Proposition 3.5.1, we find

$$|\mathtt{R}_{\Gamma,m}|^2 \leqslant C_1 \sum_{n=1}^{N_m} \Delta t^n_m \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} a_\sigma\lambda_\sigma\eta^n_\sigma |\varphi^{n-1}_K - \varphi^{n-1}_{K\sigma}|^2 \leqslant 2C_1T\|\nabla\varphi\|^2_\infty \frac{\max_i \lambda_i}{\mu} m^{d-1}(\Gamma) h_{\mathscr{T}_m},$$

so $\mathtt{R}_{\Gamma,m}\to 0$ as $m\to+\infty$. Besides, we also apply the Cauchy-Schwarz inequality to $\mathtt{R}_{i,m}$ in order to obtain

$$\begin{aligned}|\mathtt{R}_{i,m}|^2 &\leqslant C_1 \sum_{n=1}^{N_m} \Delta t^n_m \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma\,\lambda_i \big|\sqrt{\eta^n_\sigma} - \sqrt{\widetilde{\eta}^n_\sigma}\big|^2 \big|\varphi^{n-1}_K - \varphi^{n-1}_{K\sigma}\big|^2 \\ &\leqslant d\lambda_i C_1 \|\nabla\varphi\|^2_\infty \sum_{n=1}^{N_m} \Delta t^n_m \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} \left|\eta^n_\sigma - \widetilde{\eta}^n_\sigma\right|,\end{aligned}$$

where we have set

$$\widetilde{\eta}^n_\sigma = \begin{cases} \eta_i(s^n_K) & \text{if } p^n_K = p^n_{K\sigma}, \\ \left[\dfrac{\Theta_i(p^n_K) - \Theta_i(p^n_{K\sigma})}{\sqrt{\lambda_i}(p^n_K - p^n_{K\sigma})}\right]^2 & \text{otherwise.}\end{cases}$$

Define

$$\widetilde{\eta}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} \widetilde{\eta}^n_\sigma & \text{if } (t,x)\in(t^{n-1}_m, t^n_m]\times\Delta_\sigma,\ \sigma\in\bigcup_{i=1}^I \mathscr{E}_{i,m}, \\ 0 & \text{otherwise.}\end{cases}$$

Reproducing the proof of Lemma 3.6.2, we can show that

$$\widetilde{\eta}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \eta(\mathcal{S}(p,x)) \qquad \text{in } L^q(Q_T),\ 1 \leqslant q < +\infty.$$

Therefore, $\mathtt{R}_{i,m} \to 0$ as $m \to +\infty$. Putting things together in (3.6.15), we conclude that $\mathtt{B}_m^1$ and $\widetilde{\mathtt{B}}_m^1$ share the same limit, which completes the proof of Proposition 3.6.3. □

Chapter 4

Numerical strategies to solve Richards' equation in heterogeneous media

The text of this chapter is a reproduction of [19].

4.1 Introduction

The Richards equation [122] is a popular model for underground water flow in the vadose zone. It consists in a simplification of the incompressible immiscible two-phase Darcy flow model, assuming that the pressure of the gas phase is known and equal to the atmospheric pressure, see for instance [23]. Besides, Richards' equation also attracts an important interest from scientists as it provides a relatively simple model that already accounts for many difficulties occurring in complex porous media flows, like degeneracies when one phase (air or water) vanishes, or strong material heterogeneities with severe changes in the physical parameters at the interface between different rocks. We formalize mathematically in §4.1.1 the problem under consideration in this chapter, namely, Richards' equation in heterogeneous domains, before discussing on possible numerical strategies in §4.1.2.

4.1.1 The Richards equation in heterogeneous domains

Let $\Omega \subset \mathbb{R}^d$, $1 \leqslant d \leqslant 3$, be a connected open polyhedral domain, representing the porous matrix in which water flows. The porous matrix is assumed to be heterogeneous, and we particularly focus on severe variations of the rock characteristics at the interface between different rock-types. More precisely, we assume that there exist polyhedral connected and disjointed open subsets $(\Omega_i)_{1\leqslant i\leqslant I}$ such that

$$\overline{\Omega} = \bigcup_{1\leqslant i\leqslant I} \overline{\Omega}_i.$$

Each subdomain Ω_i represents a rock-type, and is assumed to be homogeneous for simplicity. We denote by

$$\Gamma_{i,j} = \overline{\Omega}_i \cap \overline{\Omega}_j, \quad 1 \leqslant i, j \leqslant I,$$

the interface between Ω_i and Ω_j and by

$$\Gamma = \bigcup_{1\leqslant i\neq j\leqslant I} \Gamma_{i,j}$$

the set containing all these interfaces.

Let $T > 0$ be an arbitrary finite time horizon, then Richards' equation in $Q_{i,T} = (0,T) \times \Omega_i$ writes

$$\phi_i \, \partial_t s + \nabla \cdot v = 0, \tag{4.1.1}$$

$$v + \lambda_i \, \eta_i(s) \, \nabla(p - \varrho g \cdot x) = 0, \tag{4.1.2}$$

$$s - \mathcal{S}_i(p) = 0. \tag{4.1.3}$$

The unknowns are the water saturation s, the water velocity v and the water pressure p. Equation (4.1.1) encodes the local conservation of the water volume (since water is described as an incompressible fluid). The Darcy-Muskat relation (4.1.2) relates the water flux to the gradient of the hydraulic head, whereas the last equation (4.1.3) links the saturation to the pressure. In the above system, ϕ_i stands for the porosity of the i-th rock and $\lambda_i > 0$ for its intrinsic permeability (isotropy of the porous medium is assumed here), while ϱ stands for the water density which is assumed to be constant, and g denotes the gravity vector. The mobility $\eta_i(s)$ is nonnegative and nondecreasing with respect to the saturation, while the function $\mathcal{S}_i$ relating the water pressure and saturation is nondecreasing and takes its values in $[0,1]$. In accordance with the classical models of the literature — see §4.4.1.3 for the precise description of the models to be used in practice in the numerical simulations — we assume that water is always mobile, i.e., that $\eta_i(\mathcal{S}_i(p)) > 0$ for all $p \in \mathbb{R}$. Water becomes immobile in the dry asymptote, i.e. $\lim_{p \to -\infty} \eta_i(\mathcal{S}_i(p)) = 0$, leading to a degeneracy of hyperbolic type. On the other hand, positive pressures correspond to saturated regimes, i.e. $\mathcal{S}_i(p) = \mathcal{S}_i(0)$ for all $p \geqslant 0$, leading to a degeneracy of elliptic type.

At the interface $\Gamma_{i,j}$, pressure and flux are continuous. More precisely, denote by p_i the trace at $(0,T) \times \Gamma_{i,j}$ of the pressure $p_{|\Omega_i}$ in $Q_{i,T}$, and by F_i the trace at $(0,T) \times \Gamma_{i,j}$ of the flux $F_{|\Omega_i}$ in $Q_{i,T}$, then the transmission conditions across $\Gamma_{i,j}$ write

$$v_i \cdot \nu_i + v_j \cdot \nu_j = 0, \tag{4.1.4}$$

$$p_i - p_j = 0, \tag{4.1.5}$$

where ν_i (resp. ν_j) denotes the normal to $\Gamma_{i,j}$ outward w.r.t. Ω_i (resp. Ω_j). Note that since the pressure is continuous, and since $\mathcal{S}_i \neq \mathcal{S}_j$ in general, the saturation is discontinuous across $\Gamma_{i,j}$. The pressure continuity (4.1.5) has to be relaxed in the case where the water mobility could vanish for finite p. We refer for instance to [27,42,65] for formulations with such relaxed pressure continuity conditions at the interfaces. Let us stress that our work can be extended without further difficulties to this more involving setting.

Concerning the boundary conditions, the external boundary $\partial\Omega$, the outward normal of which being denoted by ν, is split into a portion called Γ^{D} where a constant Dirichlet boundary condition is imposed, and $\Gamma^{\mathrm{N}} = \partial\Omega \backslash \Gamma^{\mathrm{D}}$ where a Neumann boundary condition is fixed, that is,

$$v \cdot \nu = q^{\mathrm{N}} \quad \text{on } (0,T) \times \Gamma^{\mathrm{N}}, \tag{4.1.6}$$

$$p = p^{\mathrm{D}} \quad \text{on } (0,T) \times \Gamma^{\mathrm{D}}. \tag{4.1.7}$$

Finally, the initial saturation profile is prescribed,

$$s(0,x) = s^0(x) \text{ in } \Omega. \tag{4.1.8}$$

4.1.2 Motivation and positioning of our work

Richards' equation is interesting in itself for modeling the infiltration of water in the near subsurface. This motivated the development of many numerical approaches with the aim of being robust while preserving accuracy, especially with respect to mass conservation. For numerical schemes approximating the solutions to Richards equation, we refer for instance to [49] for finite differences, to [80] for control volume finite elements, to [72,75] for two-point flux approximation (TPFA) finite volumes and to [35,100,130] for more advanced finite volume methods, to [26,120,137] for mixed finite elements, or to [106] for discontinuous Galerkin approaches. The above reference list is far from being exhaustive, and we refer to [78] for a review.

The problem being nonlinear and degenerate, an important part of the research effort has been assigned to the design of efficient iterative linearization procedures. Two main approaches then emerge: a first one based on (modified) Picard type fixed point strategies, and second one relying on Newton's method. Suitably designed Picard iteration based methods are known to enjoy robustness at the price of a mere linear convergence speed, see for instance [49,107,118,128]. On the other hand, a crude Newton's algorithm may face severe difficulties to converge, see for instance [104,107] for comparison of different approaches. This motivated the introduction of methods based on variable switch [58,81], nested Newton loops [47], or nonlinear preconditioning techniques [28] to increase robustness. Our approach, which is described in §4.2.3 and [18], relies on the so-called parametrization approach introduced in [29,34], which can be interpreted as a generalization of the variable switch approach as well as a (diagonal) nonlinear preconditioning technique.

The second main difficulty to be addressed is the strong heterogeneity of the domain Ω with discontinuous physical characteristics across Γ. Since the pressure is continuous, cf. (4.1.5), schemes that are based on formulations involving the Kirchhoff transform $\theta_i = \int_0^p \eta_i(\mathcal{S}_i(a))\mathrm{d}a$, which is known to be a powerful tool for the mathematical [9] and numerical [75,120,137] study of Richards equation, will require a specific treatment at the interfaces to maintain the continuity of the pressure. We refer for instance to [30,41,65,67,69,88] for methods built in this spirit. A more natural approach consists in using discrete fluxes expressed directly in the form (4.1.2), with degrees of freedom localized on the interface Γ to enforce the continuity of the pressure, as done for instance in [14,38, 74,90,115]. Let us also mention [3,4,127] where the authors solve the transmission condition (4.1.4)–(4.1.5) thanks to an iterative procedure stemming from domain decomposition. In the case of cell centered methods, like for instance TPFA finite volumes, convergence can also be assessed without any specific treatment of the interface, as for example done in our recent contribution [20]. However, the pressure continuity is only imposed at convergence w.r.t. grid refinement, leading to possible loss in the accuracy. Therefore, specific treatments of the interface are needed. As highlighted in [32], the specific treatment of the interface Γ may have a major impact on the Newton's method behavior. The purpose of this chapter is to compare several approaches described in §4.3 to deal with the interface transmission condition (4.1.4)–(4.1.5) and to depict their pros and cons when confronted to different physical settings described in §4.4.

4.2 Problem discretization

4.2.1 Space-time discretization

Let $(\mathscr{T},\mathscr{E})$ be a finite-volume space discretization of Ω satisfying the classical orthogonality condition required for the consistency of the Two Point Flux Approximation (TPFA), see [73, Definition 9.1] for more details. Here $\mathscr{T}$ denotes the set of cells and $\mathscr{E}$ the set of faces. We assume that the

mesh is consistent with the geometry in the sense that, for all $K \in \mathscr{T}$, there exists $i \in \{1, \dots, I\}$ such that $K \subset \Omega_i$. We denote by $\mathscr{T}_i = \{K \in \mathscr{T} \mid K \subset \Omega_i\}$. Then for all $f \in \{\mathcal{S}, \lambda, \phi, \dots\}$ that depends on the rock-type, we set $f_K = f_i$ if $K \in \mathscr{T}_i$. The set $\mathscr{E}$ is then subdivided into: the set of internal faces shared by cells of the same subdomain $\mathscr{E}_i = \{\sigma = K|L \in \mathscr{E} \mid K, L \in \mathscr{T}_i\}$, the set of the internal faces shared by cells belonging to different subdomains $\mathscr{E}_\Gamma = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma\} = \bigcup_{i \neq j} \{\sigma = K|L \in \mathscr{E} \mid K \in \mathscr{T}_i,\ L \in \mathscr{T}_j\}$, the set of Dirichlet faces $\mathscr{E}_{\text{ext}}^{\text{D}} = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \Gamma^{\text{D}}\}$ and the set of Neumann faces $\mathscr{E}_{\text{ext}}^{N} = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \Gamma^{N}\}$. Let us call $\mathscr{E}_{\text{int}} = \mathscr{E}_i \cup \mathscr{E}_\Gamma$ the set of all internal faces. We also introduce the local set $\mathscr{E}_K = \{\sigma \in \mathscr{E} \,|\, \sigma \subset \partial K\}$ containing all the faces surrounding a cell K. To each face $\sigma \in \mathscr{E}$ we associate a distance d_σ defined by

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathscr{E}_{\text{int}}, \\ d_{K,\sigma} & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\text{ext}}^{\text{D}} \cup \mathscr{E}_{\text{ext}}^{\text{N}}) \end{cases} \tag{4.2.1}$$

where, for all pair (K, σ) such that $\sigma \in \mathscr{E}_K$, $d_{K,\sigma} = |x_K - x_\sigma|$, with x_K the cell center and x_σ the face center, which is chosen as the intersection of $[x_K, x_L]$ with σ. Moreover, for each cell K, we denote by m_K its Lebesgue measure, and by m_σ the measure of a face σ. The time discretization is given by a vector of values $(t^n)_{0 \leqslant t \leqslant N}$ with $0 = t^0 < t^1 < \cdots < t^N = T$, and we denote by $t^n - t^{n-1} = \Delta t^n$, $1 \leqslant n \leqslant N$, the time-steps.

4.2.2 Upstream TPFA finite-volume scheme

The two-point flux approximation of a diffusive flux, $F_{K\sigma}$, related to the gradient of an unknown w and coming out a cell K through the face σ, is defined by

$$F_{K\sigma} = a_\sigma (w_{K\sigma} - w_K),$$

where the transmissivity on the face $\sigma \in \mathscr{E}$ is defined by

$$a_\sigma = \frac{m_\sigma}{d_\sigma}$$

and the mirror value $w_{K\sigma}$ by

$$w_{K\sigma} = \begin{cases} w_L & \text{if } \sigma = K|L \in \mathscr{E}_{\text{int}}, \\ w_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\text{ext}}^{\text{N}}, \\ w_\sigma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\text{ext}}^{\text{D}}. \end{cases}$$

The saturation capillary-pressure relationship (4.1.3) and the volume balance (4.1.1) are discretized into

$$m_K \phi_K \frac{s_K^n - s_K^{n-1}}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} F_{K\sigma}^n = 0, \qquad K \in \mathscr{T}_i,\ n \geqslant 1, \tag{4.2.2a}$$

$$s_K^n - \mathcal{S}_K(p_K^n) = 0, \qquad K \in \mathscr{T}_i, n \geqslant 1, \tag{4.2.2b}$$

where the flux

$$F_{K\sigma}^n = \begin{cases} a_\sigma \lambda_K \eta_\sigma^n \left[\vartheta_K^n - \vartheta_{K\sigma}^n\right] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\text{int}} \cup \mathscr{E}_{\text{ext}}^{\text{D}}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} \mathrm{d}t \int_\sigma q^{\text{N}} \,\mathrm{d}\gamma, & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\text{ext}}^{\text{N}}, \end{cases} \tag{4.2.3}$$

is an approximation of $\int_\sigma v^n \cdot \nu_{K,\sigma}\,\mathrm{d}\gamma$, with the hydraulic head being defined as

$$\vartheta^n = p^n + \psi = p^n - \varrho g \cdot x.$$

In (4.2.3), the face mobilities are upwinded in the following way

$$\eta_\sigma^n = \begin{cases} \eta_K(s_K^n) & \text{if } \vartheta_K^n - \vartheta_{K\sigma}^n \geqslant 0, \\ \eta_{K\sigma}(s_{K\sigma}^n) & \text{otherwise.} \end{cases} \tag{4.2.4}$$

The initial condition (4.1.8) is discretized into

$$s_K^0 = \frac{1}{m_K}\int_K s^0, \quad \forall K \in \mathscr{T}_i, \tag{4.2.5}$$

and the Dirichlet boundary condition (4.1.7) into

$$p_\sigma^{\mathrm{D}} = \frac{1}{m_\sigma}\int_\sigma p^{\mathrm{D}}, \quad \forall \sigma \in \mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}. \tag{4.2.6}$$

4.2.3 Switch of variable and parametrization technique

Let us now detail the resolution strategy for problem (4.2.2b)–(4.2.6). A natural approach to solve this nonlinear system is to choose the pressure $(p_K)_{K\in\mathscr{T}}$ as primary unknown and to solve it via an iterative method such as Newton's one. However, the pressure variable is known to be an inefficient choice for dry soils $s \ll 1$, because of the degeneracy of Richards' equation, where schemes in which saturation is the primary variable outperform. On the other hand, the knowledge of the saturation is not sufficient to describe the pressure curve in saturated regions where the pressure-saturation relation cannot be inverted. This motivated the design of schemes which introduce a switch of variable [58, 81]. Our approach is based on the technique proposed by Brenner and Cancès [29], in which a third generic variable τ is introduced to become the primary unknown of the system. Then, removing the subscript i related to the rock-type for convenience, the idea is to choose a parametrization of the graph $\{p, \mathcal{S}(p)\}$, i.e., to construct two monotone functions

$$\mathfrak{s} : I \to [s_{\mathrm{rw}}, 1 - s_{\mathrm{rn}}], \qquad \mathfrak{p} : I \to \mathbb{R},$$

such that

$$\mathfrak{s}(\tau) = \mathcal{S}(\mathfrak{p}(\tau)), \qquad 0 < \mathfrak{s}'(\tau) + \mathfrak{p}'(\tau) < \infty, \tag{4.2.7}$$

for all $\tau \in I \subset \mathbb{R}$. The latter non-degeneracy assumption ensures that for all $p \in \mathbb{R}$, there exists a unique $\tau \in \mathbb{R}$ such that $(p, \mathcal{S}(p)) = (\mathfrak{p}(\tau), \mathfrak{s}(\tau))$. Such a parametrization is not unique, for instance we can choose $I = \mathbb{R}$, $\mathfrak{p} = \mathrm{Id}$ which is equivalent to solving the system always in pressure, but this is not recommended as seen before. Here, we set $I = \mathbb{R}$ and

$$\mathfrak{s}(\tau) = \begin{cases} \mathcal{S}(\kappa(\tau - \tau_*) + p_*) & \text{if } \tau \leqslant \tau_*, \\ s_{\mathrm{rw}} + \tau(1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}) & \text{if } \tau_* \leqslant \tau \leqslant \tau_{\mathrm{s}}, \\ \mathcal{S}(p_{\mathrm{s}} + \varsigma(\tau - \tau_{\mathrm{s}})) & \text{if } \tau \geqslant \tau_{\mathrm{s}}, \end{cases} \tag{4.2.8a}$$

$$\mathfrak{p}(\tau) = \begin{cases} \kappa(\tau - \tau_*) + p_* & \text{if } \tau \leqslant \tau_*, \\ \mathcal{S}^{-1}(s_{\mathrm{rw}} + \tau(1 - s_{\mathrm{rn}} - s_{\mathrm{rw}})) & \text{if } \tau_* \leqslant \tau \leqslant \tau_{\mathrm{s}}, \\ p_{\mathrm{s}} + \varsigma(\tau - \tau_{\mathrm{s}}) & \text{if } \tau \geqslant \tau_{\mathrm{s}}. \end{cases} \tag{4.2.8b}$$

In the above formulas, $(p_{\rm s}, s_{\rm s}) = (\mathfrak{p}(\tau_{\rm s}), \mathfrak{s}(\tau_{\rm s}))$ is referred later on as the switching point, at which one passes from τ behaving as the saturation to τ behaving as the pressure (recall that Newton's iterations are not sensitive to linear changes of variables). Another switch is incorporated at $(p_*, s_*) = (\mathfrak{p}(\tau_*), \mathfrak{s}(\tau_*))$ to improve Newton's robustness in presence of heterogeneities. The parameter τ_* is chosen so small that the solution $(p_K^n, s_K^n)_{K\in\mathscr{T}}$ to the scheme is always larger that (p_*, s_*). The parameters κ and ς are chosen so that $\mathfrak{p}$ is C^1, leading to the expressions

$$\kappa = \frac{1 - s_{\rm rn} - s_{\rm rw}}{\mathcal{S}'(p_*^+)}, \quad \text{and} \quad \varsigma = \frac{1 - s_{\rm rn} - s_{\rm rw}}{\mathcal{S}'(p_{\rm s}{}^-)}, \tag{4.2.9}$$

where $\mathcal{S}'(p_*^+)$ and $\mathcal{S}'(p_{\rm s}^-)$ respectively denote the limits of $\mathcal{S}'(p)$ as p tends to p_* and $p_{\rm s}$ from above and below. Then if $\mathcal{S}$ is C^1, so is $\mathfrak{s} = \mathcal{S} \circ \mathfrak{p}$. When $\mathcal{S}$ is convex then concave, as in the Brooks-Corey and van Genuchten-Mualem settings detailed in §4.4.1.3, then choosing $\tau_{\rm s}$ such that $(p_{\rm s}, s_{\rm s})$ is the inflection point of the graph of $\mathcal{S}$ ensures that both $\mathfrak{p}$ and the restriction of $\mathfrak{s}$ to $[\tau_*, +\infty)$ are concave. Moreover, if $\mathcal{S}$ belongs to $C^2(\mathbb{R})$ as in the van Genuchten-Mualem setting, then so do the restrictions of $\mathfrak{p}$ and $\mathfrak{s}$ to $(\tau_*, +\infty)$. An example of parametrized curves $\mathfrak{p}, \mathfrak{s}$ corresponding to van Genuchten-Mualem pressure-saturation laws is shown in Figure 4.1.

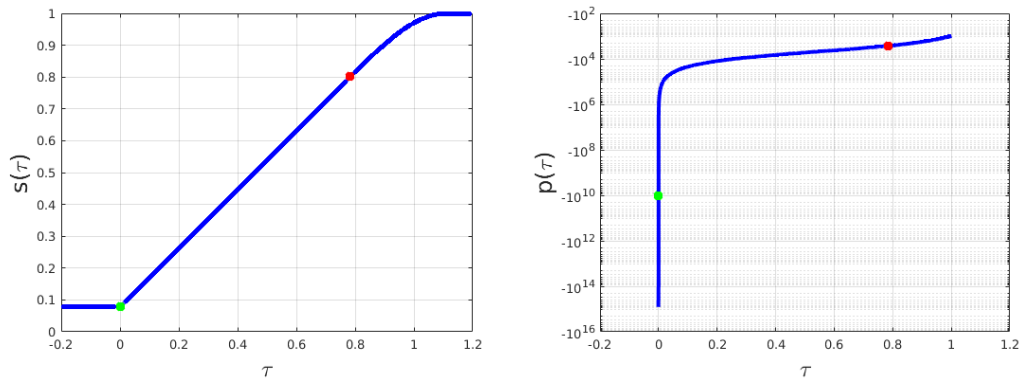


Figure 4.1: Plot of saturation and pressure parametrized van Genuchten-Mualem curves, using values of rock type 1 reported in Table 4.2 . The green dot indicate the value for $\tau = \tau_*$ and the magenta one $\tau = \tau_{\rm s}$.

Applying the parametrization to our equations, we obtain the parametrized system:

$$m_K \phi_K \frac{\mathfrak{s}_K(\tau_K^n) - \mathfrak{s}_K(\tau_K^{n-1})}{\Delta t^n} + \sum_{\sigma\in\mathscr{E}_K} F_{K\sigma}^n = 0, \qquad K \in \mathscr{T}_i,\ n \geqslant 1, \tag{4.2.10}$$

where the fluxes (4.2.3) become

$$F_{K\sigma}^n = \begin{cases} a_\sigma \lambda_K \eta_\sigma^n \left[\vartheta_K(\tau_K^n) - \vartheta_{K\sigma}(\tau_{K\sigma}^n)\right] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm D}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^n}^{t^{n+1}} {\rm d}t \int_\sigma q^{\rm N}\, {\rm d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \end{cases} \tag{4.2.11}$$

with

$$\vartheta_K(\tau) = \mathfrak{p}_K(\tau) + \psi_K = \mathfrak{p}_K(\tau) - \varrho g \cdot x_K, \tag{4.2.12}$$

and the upwinded face mobilities turn into

$$\eta_\sigma^n = \begin{cases} \eta_K(\mathfrak{s}_K(\tau_K^n)) & \text{if } \vartheta_K(\tau_K^n) - \vartheta_{K\sigma}(\tau_{K\sigma}^n) \geqslant 0, \\ \eta_{K\sigma}(\mathfrak{s}_K(\tau_{K\sigma}^n)) & \text{otherwise.} \end{cases} \tag{4.2.13}$$

Finally, we rewrite the initial condition as

$$\tau_K^0 = \mathfrak{s}_K^{-1}\left(\frac{1}{m_K}\int_K s^0\right), \quad \forall K \in \mathscr{T}_i, \tag{4.2.14}$$

and the Dirichlet boundary condition as

$$\tau_\sigma^{\mathrm{D}} = \mathfrak{p}_K^{-1}\left(\frac{1}{m_\sigma}\int_\sigma p^{\mathrm{D}}\right), \quad \forall \sigma \in \mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}. \tag{4.2.15}$$

So far, We have not specified yet how the interface fluxes $F_{K\sigma}^n$ for $\sigma = \mathscr{E}_\Gamma$ are treated. This specification is the purpose of §4.3. In the case of a homogeneous domain where $\Gamma = \varnothing$, the resulting system $\mathcal{F}_n(\boldsymbol{\tau}^n) = \mathbf{0}$ which is fully equivalent to (4.2.2b)–(4.2.6), admits a unique solution $\boldsymbol{\tau}^n$ (for details see [20, Proposition 3.6-3.7]).

Remark 4.2.1. *The practical resolution of the nonlinear system relies on Newton's method. In the homogeneous setting, the method we use is the one that is presented in [18, Section 2] with some differences. The first one concerns the approximation of the k_r law by the van Genuchten-Mualem model, that we explain in §4.4.1.3. Another one is related to the values of τ: here $I = \mathbb{R}$, so no projection of τ after each Newton iteration to avoid $\tau < s_{\mathrm{rw}}$ is required. Then, when we treat dry zones, we risk to manage a singular Jacobian matrix. In order to avoid this we impose that $k_r(s_{\mathrm{rw}}) = 10^{-33}$ when evaluating $\mathbf{J}$ to allow the pressure-gravity motor not to be zero. Finally, to help Newton's algorithm recover the good direction when the ℓ_∞ norm of the residual exceeds 10^2, a relaxation is activated with 0.3 as relaxing constant.*

4.3 Numerical treatment of the interface

This section is devoted to the presentation of different strategies to approximate the transmission conditions (4.1.4)–(4.1.5) across the faces $\sigma \in \mathscr{E}_\Gamma$ located at an interface between to different rock-types. We propose four schemes, referred to as methods A to D. For the last one, two different iterative Newton-based solvers are proposed.

4.3.1 Method A

This method basically consists in treating the interfaces as standard bulk faces, leading to the formula

$$F_{K\sigma}^n = a_\sigma \lambda_\sigma \eta_\sigma^n \left[\mathfrak{p}_K(\tau_K^n) - \mathfrak{p}_L(\tau_L^n) - \varrho g \cdot (x_K - x_L)\right], \qquad \sigma = K|L \in \mathscr{E}_\Gamma, \tag{4.3.1}$$

where the face permeabilities $(\lambda_\sigma)_{\sigma\in\mathscr{E}_\Gamma}$ are given by

$$\lambda_\sigma = \frac{\lambda_K \lambda_L d_\sigma}{\lambda_K d_{L,\sigma} + \lambda_L d_{K,\sigma}}, \qquad \sigma = K|L \in \mathscr{E}_\Gamma, \tag{4.3.2}$$

and the upwind face mobilities turn into

$$\eta_\sigma^n = \begin{cases} \eta_K(\mathfrak{s}_K(\tau_K^n)) & \text{if } \mathfrak{p}_K(\tau_K^n) - \mathfrak{p}_L(\tau_L^n) - \varrho g \cdot (x_K - x_L) \geqslant 0, \\ \eta_L(\mathfrak{s}_L(\tau_L^n)) & \text{otherwise.} \end{cases} \tag{4.3.3}$$

Therefore, the continuity of the normal flux (4.1.4) is exactly transposed into the local conservation condition

$$F_{K\sigma}^n + F_{L\sigma}^n = 0, \qquad \sigma = K|L \in \mathscr{E}_\Gamma,\ n \geqslant 1. \tag{4.3.4}$$

On the other hand, $p_K^n \neq p_L^n$ in general. The pressure continuity (4.1.5) is recovered asymptotically as d_σ tends to 0 from (4.3.1). More precisely, assuming that $|F_{K\sigma}^n| \leqslant C m_\sigma$, then we deduce from (4.3.1) that $|\mathfrak{p}_K(\tau_K^n) - \mathfrak{p}_L(\tau_L^n)| \leqslant C d_\sigma$, where the constant C has been updated and further depends on $\lambda_\sigma, \max_i \|\eta_i \circ \mathcal{S}_i\|_\infty$ and ϱg.

The scheme (4.2.10)–(4.2.13), complemented by the interface fluxes (4.3.1)–(4.3.3), has been shown in [20] to be well-posed in the sense that the corresponding nonlinear system admits a unique solution $(\tau_K^n)_{K\in\mathscr{T}}$. Further, the rigorous convergence of the scheme as the mesh size and the time-steps tend to 0 is also established. However, the numerical results presented in [20] (as well as those presented in what follows) show that the expected first order convergence can be lost in presence of heterogeneities. Methods B, C, and D have been designed as remedies to this loss of accuracy, which takes its origin in the poor approximation of the pressure continuity (4.1.5) by Method A.

4.3.2 Method B

This method, introduced in [20] consists in adding two thin cells, denoted by $I_{\sigma,K}$ and $I_{\sigma,L}$, of thickness $\delta_B \ll d_\sigma$ on both sides of each face $\sigma = K|L \in \mathscr{E}_\Gamma$ located at a rock-type interface, as depicted in Figure 4.2. This leads to the adjunction of two additional unknowns $\tau_{\sigma,K}^n$ and $\tau_{\sigma,L}^n$ per interface $\sigma = K|L \in \mathscr{E}_\Gamma$, that will allow for a more precise approximation of the pressure continuity condition (4.1.5).

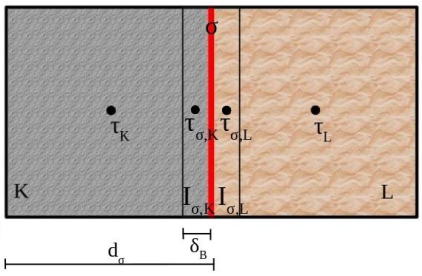


Figure 4.2: Method B: introduction of two thin cells on both sides of a face located between two rock types.

Define $(F_{K\sigma}^n)_{\sigma\in\mathscr{E}_\Gamma}$ to be used in (4.2.10) by setting, for $\sigma = K|L \in \mathscr{E}_\Gamma$ and $n \geqslant 1$,

$$F_{K\sigma}^n = a_{\sigma,K}\lambda_K \eta_{\sigma,K}^n \left[\mathfrak{p}_K(\tau_K^n) - \mathfrak{p}_K(\tau_{\sigma,K}^n) - \varrho g \cdot (x_K - x_{\sigma,K})\right], \tag{4.3.5a}$$

$$F_{L\sigma}^n = \ a_{\sigma,L}\lambda_L \eta_{\sigma,L}^n \left[\mathfrak{p}_L(\tau_L^n) - \mathfrak{p}_L(\tau_{\sigma,L}^n) - \varrho g \cdot (x_L - x_{\sigma,L})\right], \tag{4.3.5b}$$

where we have set

$$a_{\sigma,K} = \frac{m_\sigma}{d_{K,\sigma} - \delta_{\mathrm{B}}/2}, \qquad x_{\sigma,K} = x_K + \frac{d_{K,\sigma} - \delta_{\mathrm{B}}/2}{d_\sigma}(x_L - x_K), \tag{4.3.6a}$$

$$a_{\sigma,L} = \frac{m_\sigma}{d_{L,\sigma} - \delta_{\mathrm{B}}/2}, \qquad x_{\sigma,L} = x_L + \frac{d_{L,\sigma} - \delta_{\mathrm{B}}/2}{d_\sigma}(x_K - x_L), \tag{4.3.6b}$$

and

$$\eta^n_{\sigma,K} = \begin{cases} \eta_K(\mathfrak{s}_K(\tau^n_K)) & \text{if } \mathfrak{p}_K(\tau^n_K) - \varrho g \cdot x_K \geqslant \mathfrak{p}_K(\tau^n_{\sigma,K}) - \varrho g \cdot x_{\sigma,K}, \\ \eta_K(\mathfrak{s}_K(\tau^n_{\sigma,K})) & \text{otherwise.} \end{cases}$$

Two equations are required to determine $\tau^n_{\sigma,K}$ and $\tau^n_{\sigma,L}$. These equations are local conservation laws in the thin cells $I_{\sigma,K}$ and $I_{\sigma,L}$. Denote by $m_{\sigma,K}$ and $m_{\sigma,L}$ the Lebesgue measure of the thin cells $I_{\sigma,K}$ and $I_{\sigma,L}$ respectively ($m_{K\sigma} = m_\sigma \delta_{\mathrm{B}}$ for Cartesian grids as depicted in Figure 4.2), then $(\tau^n_{K\sigma}, \tau^n_{L\sigma})$ are determined by

$$m_{\sigma,K}\phi_K \frac{\mathfrak{s}_K(\tau^n_{\sigma,K}) - \mathfrak{s}_K(\tau^{n-1}_{\sigma,K})}{\Delta t} + F^n_\sigma - F^n_{K\sigma} = 0, \tag{4.3.7a}$$

$$m_{\sigma,L}\phi_L \frac{\mathfrak{s}_L(\tau^n_{\sigma,L}) - \mathfrak{s}_L(\tau^{n-1}_{\sigma,L})}{\Delta t} - F^n_\sigma - F^n_{L\sigma} = 0, \tag{4.3.7b}$$

where F^n_σ is the flux from $I_{\sigma,K}$ to $I_{\sigma,L}$ defined by

$$F^n_\sigma = \frac{m_\sigma}{\delta_{\mathrm{B}}} \lambda_\sigma \eta^n_\sigma \left[\mathfrak{p}_K(\tau^n_{\sigma,K}) - \mathfrak{p}_L(\tau^n_{\sigma,L}) - \varrho g \cdot (x_{\sigma,K} - x_{\sigma,L})\right], \tag{4.3.8}$$

with λ_σ given by (4.3.2) and

$$\eta^n_\sigma = \begin{cases} \eta_K(\mathfrak{s}_K(\tau^n_{\sigma,K})) & \text{if } \mathfrak{p}_K(\tau^n_{\sigma,K}) - \varrho g \cdot x_{\sigma,K} \geqslant \mathfrak{p}_L(\tau^n_{\sigma,L}) - \varrho g \cdot x_{\sigma,L}, \\ \eta_L(\mathfrak{s}_L(\tau^n_{\sigma,L})) & \text{otherwise.} \end{cases} \tag{4.3.9}$$

Assuming that $|F^n_\sigma| \leqslant C m_\sigma$, then we deduce from (4.3.8) that $|\mathfrak{p}_K(\tau^n_{\sigma,K}) - \mathfrak{p}_L(\tau^n_{\sigma,L})| \leqslant C\delta_{\mathrm{B}}$, improving the pressure continuity with respect to Method A since $\delta_{\mathrm{B}} \ll d_\sigma$. On the other hand, summing (4.3.7a) and (4.3.7b) yields

$$|F^n_{K\sigma} + F^n_{L\sigma}| \leqslant C \frac{m_\sigma \delta_{\mathrm{B}}}{\Delta t} \underset{\delta_{\mathrm{B}} \to 0}{\longrightarrow} 0. \tag{4.3.10}$$

Note that even if (4.3.10) can be interpreted as a defect in the approximation of (4.1.4), Method B is still conservative since we keep track of this defect thanks to the discrete conservation laws (4.3.7a)–(4.3.7b). The volume m_K of the cell K is updated into

$$m_K \leftarrow m_K - \sum_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_\Gamma} m_{\sigma,K} \tag{4.3.11}$$

in (4.2.10) for all K having interface edges.

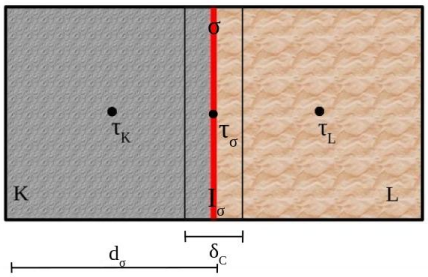


Figure 4.3: Method C: one extra thin cell I_σ overlaps the interface located between two rock types.

4.3.3 Method C

This method takes inspiration from [32,34] and consists in adding only one thin cell, I_σ, of thickness $\delta_C \ll d_\sigma$, which overlaps the rock-type interface as shown in Figure 4.3.

For $\sigma = K|L \in \mathscr{E}_\Gamma$, we denote by $m_{\sigma,K}$ and $m_{\sigma,L}$ the Lebesgue measures of $I_{\sigma,K} := I_\sigma \cap K$ and $I_{\sigma,L} := I_\sigma \cap L$ respectively. The system is enriched with only one extra unknown τ_σ^n per face $\sigma \in \mathscr{E}_\Gamma$, in opposition to Method B where two additional unknowns where needed. The new cell I_σ is shared by two subcells $I_{\sigma,K}$ and $I_{\sigma,L}$ corresponding to different lithologies. To enforce one single pressure in the cell, we introduce a second parametrization and define monotone functions $\omega_{\sigma,K}, \omega_{\sigma,L}$ with $\omega'_{\sigma,K} + \omega'_{\sigma,L} > 0$ such that

$$\mathfrak{p}_K(\omega_{\sigma,K}(\tau)) = \mathfrak{p}_L(\omega_{\sigma,L}(\tau)), \qquad \forall \tau. \tag{4.3.12}$$

As for the parametrization $(\mathfrak{p}, \mathfrak{s})$ of the graph of $\mathcal{S}$, an infinite number of admissible $(\omega_{\sigma,K}, \omega_{\sigma,L})$ satisfying (4.3.12) can be built. We further investigate two choices.

The first possibility, named with exponent 1, consists in setting

$$\omega^1_{\sigma,K}(\tau) = \tau, \quad \omega^1_{\sigma,L}(\tau) = \mathfrak{p}_L^{-1} \circ \mathfrak{p}_K(\tau), \tag{4.3.13}$$

the orientation of the cell being such that $\omega^1_{\sigma,L}$ is concave (choose K and L such that $p_{b,K} > p_{b,L}$ or $\xi_K > \xi_L$ in the Brooks-Corey and van Genuchten-Mualem settings described in §4.4.1.3 respectively), hence so does $\mathfrak{p}_L \circ \omega^1_{\sigma,L}$. As it appears on figures 4.4 and 4.5, the derivative of $\omega^1_{\sigma,L}$ might blow up (the value of κ_K defined by (4.2.9) is very large in practice).

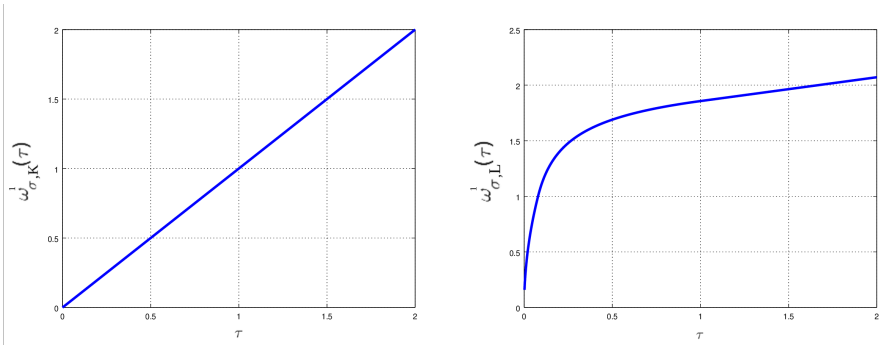


Figure 4.4: Behaviour of $\omega^1_{\sigma,K}(\cdot)$ and $\omega^1_{\sigma,L}(\cdot)$ functions using the Brooks-Corey model.

Our second proposition, named with exponent 2, is tailored to maintain control on the derivatives of $\omega^2_{\sigma,K}$ and $\omega^2_{\sigma,L}$. To this end, keeping the same orientation $K|L$ of the interface σ, we set

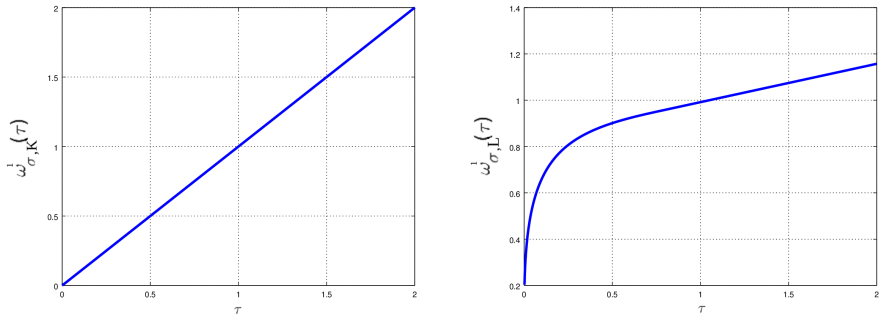


Figure 4.5: Behaviour of $\omega^1_{\sigma,K}(\cdot)$ and $\omega^1_{\sigma,L}(\cdot)$ functions using the van Genuchten-Mualem model.

$$\omega^2_{\sigma,K}(\tau) = \begin{cases} \mathfrak{p}_K^{-1} \circ \mathfrak{p}_L(\tau) & \text{if } \tau \leqslant \beta_L, \\ \tau + \beta_K - \beta_L & \text{if } \tau \geqslant \beta_L. \end{cases} \tag{4.3.14a}$$

$$\omega^2_{\sigma,L}(\tau) = \begin{cases} \tau & \text{if } \tau \leqslant \beta_L, \\ \mathfrak{p}_L^{-1} \circ \mathfrak{p}_K(\tau + \beta_K - \beta_L) & \text{if } \tau \geqslant \beta_L. \end{cases} \tag{4.3.14b}$$

the parameters β_K and β_L are uniquely determined by the conditions

$$\mathfrak{p}_K(\beta_K) = \mathfrak{p}_L(\beta_L), \qquad \mathfrak{p}'_K(\beta_K) = \mathfrak{p}'_L(\beta_L),$$

since $\mathfrak{p}_K, \mathfrak{p}_L$ are increasing and concave. This yields 1-Lipschitz continuous functions $\omega^2_{\sigma,K}$ and $\omega^2_{\sigma,L}$ as depicted on Figures 4.6 and 4.7.

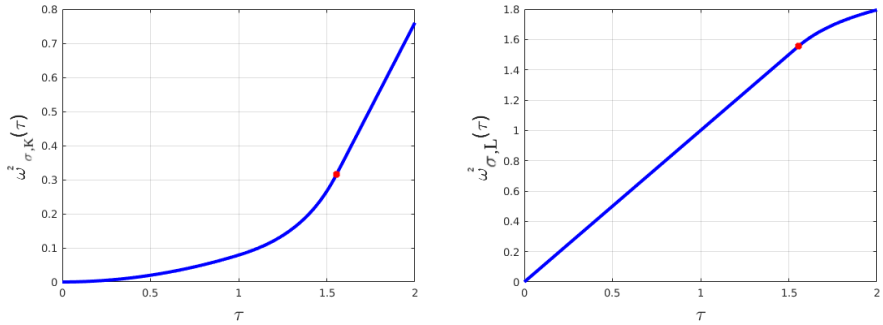


Figure 4.6: Behaviour of $\omega^2_{\sigma,K}(\cdot)$ and $\omega^2_{\sigma,L}(\cdot)$ functions using the Brooks-Corey model.

With a double parametrization $(\omega_{\sigma,K}, \omega_{\sigma,L})_{\sigma\in\mathscr{E}_\Gamma}$ at hand, Method C then consists in writing a discrete conservation law in I_σ. While in Method B, the sub-cells $I_{\sigma,K}$ and $I_{\sigma,L}$ had different pressures generating an in-between flux F^n_σ (4.3.8), here the two sub-cells share the same pressure

$$p^n_\sigma = \mathfrak{p}_K(\omega_{\sigma,K}(\tau^n_\sigma)) = \mathfrak{p}_L(\omega_{\sigma,L}(\tau^n_\sigma)), \qquad \sigma = K|L \in \mathscr{E}_\Gamma, \tag{4.3.15}$$

thanks to (4.3.12). The discrete volume conservation on I_σ then reads

$$\begin{aligned} m_{\sigma,K}\phi_K \frac{\mathfrak{s}_K(\omega_{\sigma,K}(\tau^n_\sigma)) - \mathfrak{s}_K(\omega_{\sigma,K}(\tau^{n-1}_\sigma))}{\Delta t^n} + m_{\sigma,L}\phi_L \frac{\mathfrak{s}_L(\omega_{\sigma,L}(\tau^n_\sigma)) - \mathfrak{s}_L(\omega_{\sigma,L}(\tau^{n-1}_\sigma))}{\Delta t^n} \\ - F^n_{K\sigma} - F^n_{L\sigma} = 0, \end{aligned} \tag{4.3.16}$$

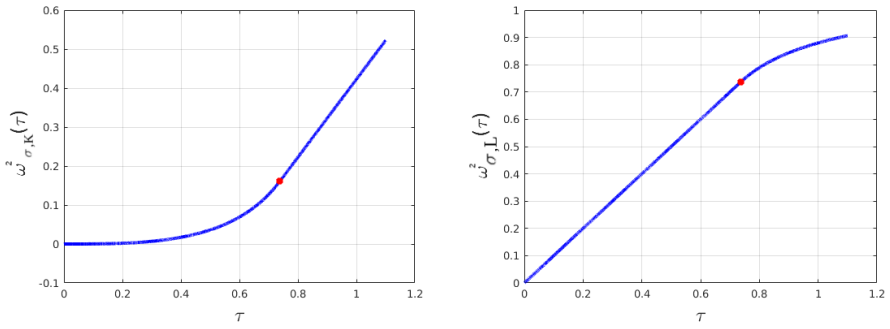


Figure 4.7: Behaviour of $\omega^2_{\sigma,K}(\cdot)$ and $\omega^2_{\sigma,L}(\cdot)$ functions using the van Genuchten-Mualem model.

where the fluxes $F^n_{K\sigma}$ and $F^n_{L\sigma}$ from K to I_σ and from L to I_σ are given by

$$F^n_{K\sigma} = \frac{m_\sigma}{d_{K,\sigma}} \lambda_K \eta^n_{\sigma,K} \left[\mathfrak{p}_K(\tau^n_K) - p^n_\sigma + \varrho g \cdot (x_K - x_\sigma)\right], \tag{4.3.17a}$$

$$F^n_{L\sigma} = \frac{m_\sigma}{d_{L,\sigma}} \lambda_L \eta^n_{\sigma,L} \left[\mathfrak{p}_L(\tau^n_L) - p^n_\sigma + \varrho g \cdot (x_L - x_\sigma)\right], \tag{4.3.17b}$$

with

$$\eta^n_{\sigma,K} = \begin{cases} \eta_K(\mathfrak{s}_K(\tau^n_K)) & \text{if } \mathfrak{p}_K(\tau^n_K) - \varrho g \cdot x_K \geqslant p^n_\sigma - \varrho g \cdot x_\sigma, \\ \eta_K(\mathfrak{s}_K(\omega_{\sigma,K}(\tau^n_\sigma))) & \text{otherwise.} \end{cases} \tag{4.3.18}$$

In (4.3.17)–(4.3.18), p^n_σ is given by (4.3.15), which should be thought as the discrete counterpart to the pressure continuity (4.1.5) across the interface. Concerning the continuity of the fluxes (4.1.4), it follows from (4.3.16) that

$$|F^n_{K\sigma} + F^n_{L\sigma}| \leqslant C \frac{m_\sigma \delta_{\mathrm{C}}}{\Delta t} \underset{\delta_{\mathrm{C}} \to 0}{\longrightarrow} 0, \tag{4.3.19}$$

meaning that (4.1.4) is recovered only asymptotically. Nevertheless, with δ_{C} small, flux continuity is captured in an accurate way. Moreover, as for Method B, Method C is locally conservative if one corrects the cell size m_K as prescribed by (4.3.11).

4.3.4 Method D

The last method we propose, referred to as Method D, consists in enforcing both the pressure continuity and the flux continuity across the interface, at the price of one edge unknown τ^n_σ on each $\sigma \in \mathscr{E}_\Gamma$ on the interface between different rocks. Such an approach has already been proposed for instance in [30, 32, 41, 65]. Letting δ_{C} tend to 0 in Method C (cf. Figure 4.8, and more precisely in (4.3.16)), one recovers the flux continuity

$$F^n_{K\sigma} + F^n_{L\sigma} = 0, \qquad \sigma = K|L \in \mathscr{E}_\Gamma, \tag{4.3.20}$$

with $F^n_{K\sigma}$ and $F^n_{L\sigma}$ respectively defined by (4.3.17) and (4.3.17b), whereas pressure continuity is still ensured by (4.3.15). We propose then two numerical strategies, later referred to as Methods D_1 and D_2 to solve the resulting nonlinear system.

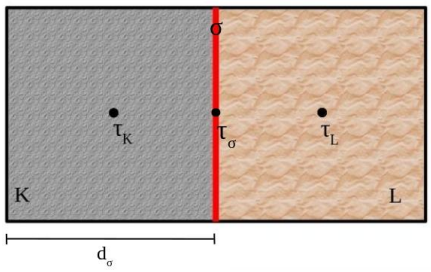


Figure 4.8: Method D: introduction of a face unknown τ_σ^n with no associated volume.

4.3.4.1 Method D_1: Schur complement based elimination of the face unknowns

With the rock-type face unknowns the obtained system is made of $\#\mathscr{T} + \#\mathscr{E}_\Gamma$ equations

$$\mathcal{F}(\boldsymbol{\tau}_{\mathscr{T}}^n, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n) = \begin{bmatrix} \mathcal{F}_{\mathscr{T}}(\boldsymbol{\tau}_{\mathscr{T}}^n, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n) \\ \mathcal{F}_{\mathscr{E}_\Gamma}(\boldsymbol{\tau}_{\mathscr{T}}^n, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n) \end{bmatrix} = \mathbf{0} \tag{4.3.21}$$

where $\boldsymbol{\tau}_{\mathscr{T}}^n = (\tau_K^n)_{K\in\mathscr{T}}$, $\boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n = (\tau_\sigma^n)_{\sigma\in\mathscr{E}_\Gamma}$, and where $\mathcal{F}_{\mathscr{T}}$ corresponds to the volume conservation laws (4.2.10) and $\mathcal{F}_{\mathscr{E}_\Gamma}$ to the flux conservation across the interfaces (4.3.20).

In what follows, we are interested in the resolution of the system (4.3.21) at a prescribed time-step n. For notation convenience, the superscript n is dropped in this section. Denote by $(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell)_{\ell\geqslant 0}$ a sequence of approximation of $(\boldsymbol{\tau}_{\mathscr{T}}, \boldsymbol{\tau}_{\mathscr{E}_\Gamma})$ given by iterations of Newton's method. The Jacobian matrix of $\mathcal{F}$ at $(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell)$, $\ell \geqslant 0$, can be split into four blocks as

$$\mathbf{J}_{\mathcal{F}}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell) = \begin{bmatrix} \mathbf{A}^\ell & \mathbf{B}^\ell \\ \mathbf{C}^\ell & \mathbf{D}^\ell \end{bmatrix}, \tag{4.3.22}$$

where

$$\mathbf{A}^\ell = \frac{\partial \mathcal{F}_{\mathscr{T}}}{\partial \boldsymbol{\tau}_{\mathscr{T}}}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell), \qquad \mathbf{B}^\ell = \frac{\partial \mathcal{F}_{\mathscr{T}}}{\partial \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell), \tag{4.3.23a}$$

$$\mathbf{C}^\ell = \frac{\partial \mathcal{F}_{\mathscr{E}_\Gamma}}{\partial \boldsymbol{\tau}_{\mathscr{T}}^n}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell), \qquad \mathbf{D}^\ell = \frac{\partial \mathcal{F}_{\mathscr{E}_\Gamma}}{\partial \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^n}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell). \tag{4.3.23b}$$

Then the matrix $\mathbf{D}^\ell$ is diagonal with negative diagonal entries because of the monotonicity of $F_{K\sigma}^n$, $F_{L\sigma}^n$ with respect to τ_σ^n that can be deduced from the monotonicity of $\mathfrak{p}_K$, $\mathfrak{p}_L$ and $\omega_{\sigma,K}$, $\omega_{\sigma,L}$. Therefore, $\mathbf{D}^\ell$ can be inverted for free.

A Newton iteration to solve (4.3.21) then computes an increment $\boldsymbol{\delta}^\ell = (\boldsymbol{\delta}_{\mathscr{T}}^\ell, \boldsymbol{\delta}_{\mathscr{E}_\Gamma}^\ell)^T$ that solves

$$\mathbf{J}_{\mathcal{F}}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell)\boldsymbol{\delta}^\ell = -\mathcal{F}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell) = -\begin{bmatrix} \mathcal{F}_{\mathscr{T}}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell) \\ \mathcal{F}_{\mathscr{E}_\Gamma}(\boldsymbol{\tau}_{\mathscr{T}}^\ell, \boldsymbol{\tau}_{\mathscr{E}_\Gamma}^\ell) \end{bmatrix} = -\begin{bmatrix} \mathcal{F}_{\mathscr{T}}^\ell \\ \mathcal{F}_{\mathscr{E}_\Gamma}^\ell \end{bmatrix},$$

or equivalently

$$(\mathbf{A}^\ell - \mathbf{B}^\ell(\mathbf{D}^\ell)^{-1}\mathbf{C}^\ell)\boldsymbol{\delta}_{\mathscr{T}}^\ell = -\mathcal{F}_{\mathscr{T}}^\ell + \mathbf{B}^\ell(\mathbf{D}^\ell)^{-1}\mathcal{F}_{\mathscr{E}_\Gamma}^\ell, \tag{4.3.24a}$$

$$\mathbf{D}^\ell\boldsymbol{\delta}_{\mathscr{E}_\Gamma}^\ell = -\mathcal{F}_{\mathscr{E}_\Gamma}^\ell - \mathbf{C}^\ell\boldsymbol{\delta}_{\mathscr{T}}^\ell. \tag{4.3.24b}$$

After solving the linear system (4.3.24a) of size $\#\mathscr{T} \times \#\mathscr{T}$ for $\boldsymbol{\delta}_{\mathscr{T}}^\ell$, inferring $\boldsymbol{\delta}_{\mathscr{E}_\Gamma}^\ell$ by (4.3.24b), the unknowns are updated by $\boldsymbol{\tau}^{\ell+1} = \boldsymbol{\tau}^\ell + \boldsymbol{\delta}^\ell$.

4.3.4.2 Method D_2: face unknowns elimination thanks to a bisection method

We present here an alternative approach to solve the nonlinear system (4.3.21). The strategy consists here in computing increments of the cell unknowns via Newton's method and updating the face unknowns by solving exactly the flux conservation (4.3.20) on each interface and at each Newton iteration ℓ. More specifically, instead of (4.3.24), the algorithm reads

$$(\mathbf{A}^\ell - \mathbf{B}^\ell(\mathbf{D}^\ell)^{-1}\mathbf{C}^\ell)\boldsymbol{\delta}^\ell_{\mathscr{T}} = -\mathcal{F}^\ell_{\mathscr{T}} + \mathbf{B}^\ell(\mathbf{D}^\ell)^{-1}\mathcal{F}^\ell_{\mathscr{E}_\Gamma}, \tag{4.3.25a}$$

$$\boldsymbol{\tau}^{\ell+1}_{\mathscr{T}} = \boldsymbol{\tau}^\ell_{\mathscr{T}} + \boldsymbol{\delta}^\ell_{\mathscr{T}}, \tag{4.3.25b}$$

$$\mathcal{F}_{\mathscr{E}_\Gamma}(\boldsymbol{\tau}^{\ell+1}_{\mathscr{T}}, \boldsymbol{\tau}^{\ell+1}_{\mathscr{E}_\Gamma}) = 0. \tag{4.3.25c}$$

In the last step, we solve the nonlinear equation (4.3.25c) for $\boldsymbol{\tau}^{\ell+1}_{\mathscr{E}_\Gamma}$ with a known value of $\boldsymbol{\tau}^{\ell+1}_{\mathscr{T}}$. This can be achieved with the help of a bisection method.

We now further detail how to solve (4.3.25c) knowing $(\boldsymbol{\tau}^\ell_K)_{K\in\mathscr{T}}$. For each $\sigma \in \mathscr{E}_\Gamma$ and for all outer Newton loop iteration ℓ, we build a sequence

$$(\vartheta^{\ell,k}_\sigma)_{k\geqslant 0} = (p^{\ell,k}_\sigma - \varrho g\cdot x_\sigma)_{k\geqslant 0}$$

approximating the interface hydraulic head at the interface. More precisely, define

$$\vartheta^\ell_K = \mathfrak{p}_K(\tau^\ell_K) - \varrho g\cdot x_K \tag{4.3.26}$$

for $K\in\mathscr{T}$ and

$$F^\ell_{K,\sigma}(\vartheta_\sigma) = \frac{m_\sigma}{d_{K,\sigma}}\lambda_K\eta^\ell_{\sigma,K}(\vartheta_\sigma)[\vartheta^\ell_K - \vartheta_\sigma], \tag{4.3.27a}$$

$$F^\ell_{L,\sigma}(\vartheta_\sigma) = \frac{m_\sigma}{d_{L,\sigma}}\lambda_L\eta^\ell_{\sigma,L}(\vartheta_\sigma)[\vartheta^\ell_L - \vartheta_\sigma], \tag{4.3.27b}$$

for $\sigma = K|L\in\mathscr{E}_\Gamma$, with

$$\eta^\ell_{\sigma,K}(\vartheta_\sigma) = \begin{cases} \eta_K\circ\mathfrak{s}_K\circ\mathfrak{p}_K^{-1}(\vartheta^\ell_K + \varrho g\cdot x_K) & \text{if } \vartheta_\sigma\leqslant\vartheta^\ell_K,\\ \eta_K\circ\mathfrak{s}_K\circ\mathfrak{p}_K^{-1}(\vartheta_\sigma + \varrho g\cdot x_\sigma) & \text{otherwise}\end{cases} \tag{4.3.28}$$

and a similar definition for $\eta^\ell_{\sigma,L}$. Then, one readily checks that $F^\ell_{K,\sigma}$ is decreasing w.r.t. ϑ_σ, and that

$$F^\ell_{K,\sigma}(\min(\vartheta^\ell_K,\vartheta^\ell_L)) \geqslant 0 \geqslant F^\ell_{K,\sigma}(\max(\vartheta^\ell_K,\vartheta^\ell_L)).$$

Therefore, the continuous and decreasing function $G^\ell_\sigma:\mathbb{R}\to\mathbb{R}$ defined by

$$G^\ell_\sigma(\vartheta_\sigma) = \frac{F^\ell_{K,\sigma}(\vartheta_\sigma) + F^\ell_{L,\sigma}(\vartheta_\sigma)}{a_\sigma\mu^{-1}(\lambda_K+\lambda_L)(|\vartheta^\ell_K| + |\vartheta^\ell_L|)}$$

vanishes at some $\vartheta^\ell_\sigma\in\big[\min(\vartheta^\ell_K,\vartheta^\ell_L),\max(\vartheta^\ell_K,\vartheta^\ell_L)\big]$, from which one deduces

$$\tau^\ell_\sigma = \omega^{-1}_{\sigma,K}\circ\mathfrak{p}_K^{-1}(\vartheta^\ell_\sigma + \varrho g\cdot x_\sigma) = \omega^{-1}_{\sigma,L}\circ\mathfrak{p}_L^{-1}(\vartheta^\ell_\sigma + \varrho g x_\sigma)$$

with $(\omega_{\sigma,K},\omega_{\sigma,L})$ being a double parametrization as introduced in §4.3.3. Then we solve the nonlinear equation $G^\ell_\sigma(\vartheta_\sigma)=0$ thanks to the classical bisection method, stopping the iterations over k when either $|G^\ell_\sigma(\vartheta^k_\sigma)|<\epsilon_{\rm bis}$ or $|\vartheta^k_0-\vartheta^k_1|<\gamma\min(|\vartheta^k_0|,|\vartheta^k_1|)$ with $\epsilon_{\rm bis}=10^{-16}$ and $\gamma=10^{-15}$ in our simulations. We can finally remark that Method D_2 does not depend on the choice of the double parametrization.

4.4 Numerical results

We now present numerical results obtained for different test cases. In all cases, we consider a two-dimensional layered domain $\Omega = [0, 5] \times [-3, 0]$ (in meters) made up of two rock types denoted by RT0 and RT1 respectively, RT1 being less permeable than RT0. The domain Ω is partitioned into three connected subdomains: $\Omega_1 = [1, 4] \times [-1, 0]$, $\Omega_2 = [0, 5] \times [-3, -2]$ (in meters) and $\Omega_3 = \Omega \backslash (\Omega_1 \cup \Omega_2)$, as depicted in Figure 4.9.

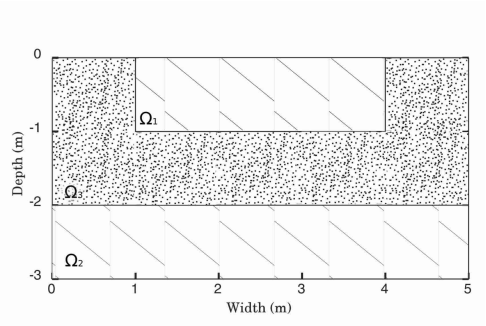


Figure 4.9: Simulation domain $\Omega = [0\mathrm{m}, 5\mathrm{m}] \times [-3\mathrm{m}, 0\mathrm{m}]$.

4.4.1 Description of the test cases

Both filling and drainage configurations are considered along with the two classical Brooks-Corey and van Genuchten-Mualem hydraulic models. These analytical models are first used in a setting where the pressure-saturation relationship and its inverse have moderate derivatives (non-steep cases). We then only consider the Brooks-Corey model and coefficients where the pressure-saturation dependence has sharp variations (steep cases).

4.4.1.1 Filling case

This test case has already been considered in [47, 81, 99, 109]. The rock-type repartition is reported in Figure 4.10. Starting from an initially dry domain Ω, where the initial capillary pressure is set to $-47.088 \cdot 10^5$ Pa, water flows from a portion $\Gamma^{\mathrm{N}} = \{(x, y) \,|\, x \in [1, 4], y = 0\}$ (in meters) of the top boundary at a constant rate of 0.5 m/day. A no-flow boundary condition is applied elsewhere. The simulation stops after 1 day.

Water flows according to the following dynamics. It starts invading the dry porous space in Ω_1. When it reaches the interface with Ω_3, capillary forces create a suction force on water from Ω_1 to Ω_3. But, on the other hand, the low permeability value in RT1 is set against this water flow through Ω_3. The simulation ends before water reaches the bottom part corresponding to Ω_2.

4.4.1.2 Drainage case

This test case is designed as a two-dimensional extension of a one-dimensional test case proposed by [108] and addressed in [47,109]. We simulate a vertical drainage starting from saturated initial and boundary conditions during $105 \cdot 10^4$ s. The initial pressure is hydro-static, that is $p^0(z) = -\rho g z$. A Dirichlet boundary condition $p^{\mathrm{D}} = 0$ Pa is imposed on the bottom boundary, $\Gamma^{\mathrm{D}} = \{(x, y) \,|\, x \in [0, 5], y = -3\}$ (in meters). The rock-type distribution of Ω is shown in Figure 4.11 along with the bottom boundary condition.

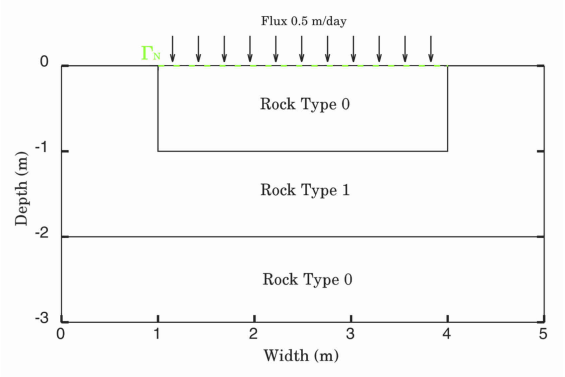


Figure 4.10: Boundary condition for the filling case.

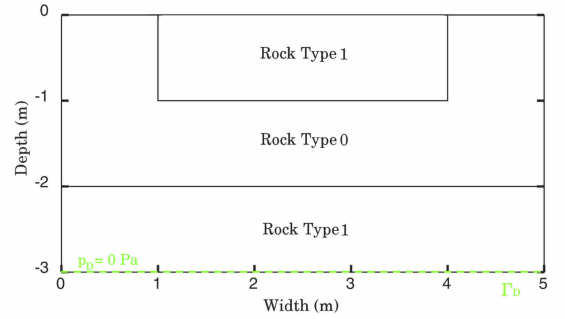


Figure 4.11: Boundary condition for the drainage test.

Note that rock types RT0 and RT1 are here reversed compared to the previous case. Thus, at the top interface between Ω_1 and Ω_3, capillarity acts in opposition to gravity and to the evolution of the system towards a dryer configuration. The interface between Ω_2 and Ω_3 acts in the opposite way: both capillarity and gravity contribute to the drainage of the RT0 subdomain.

4.4.1.3 Hydraulic models

For two-phase problems, water saturation and capillary pressure are linked through the relation $s = \mathcal{S}(p)$. Here $\mathcal{S} : \mathbb{R} \to [0,1]$ is nondecreasing. It satisfies $\mathcal{S}(p) = 1 - s_{\mathrm{rn}}$ if $p \geqslant p_b$ and $\mathcal{S}(p) \to s_{\mathrm{rw}}$ as $p \to -\infty$, with s_{rw} (resp. s_{rn}) the residual wetting (resp. non-wetting) saturation. In the following, we define the effective saturation

$$s_{\mathrm{eff}} = \Pi_{[0,1]}\left(\frac{s - s_{\mathrm{rw}}}{(1 - s_{\mathrm{rn}}) - s_{\mathrm{rw}}}\right), \tag{4.4.1}$$

where $\Pi_{[0,1]}$ is the projection on $[0,1]$. To model the two-phase flow characteristics for both rock types, we use

- either the Brooks-Corey [39] model:

$$k_r(s) = s_{\text{eff}}^{3+2/n}, \tag{4.4.2a}$$

$$\mathcal{S}(p) = \begin{cases} s_{\text{rw}} + (1 - s_{\text{rn}} - s_{\text{rw}})\left(-\dfrac{p}{p_b}\right)^{-n} & \text{if } p \leqslant -p_b, \\ 1 - s_{\text{rn}} & \text{if } p > -p_b, \end{cases} \tag{4.4.2b}$$

- or the van Genuchten-Mualem [133] model:

$$k_r(s) = s_{\text{eff}}^{1/2}\big\{1 - \big[1 - s_{\text{eff}}^{1/m}\big]^m\big\}^2, \tag{4.4.3a}$$

$$\mathcal{S}(p) = \begin{cases} s_{\text{rw}} + (1 - s_{\text{rn}} - s_{\text{rw}})\left[1 + \left|\dfrac{\xi p}{\varrho g}\right|^n\right]^{-m} & \text{if } p \leqslant 0, \\ 1 - s_{\text{rn}} & \text{if } p > 0, \end{cases} \tag{4.4.3b}$$

  with $m = 1 - 1/n$.

In both models, we have denoted by $k_r(\cdot)$ the relative permeability which, with the water viscosity $\mu = 10^{-3}$ Pa $\cdot$ s, defines the water mobility thanks to $\eta(\cdot) = k_r(\cdot)/\mu$. The parameters used for both rock types are given in Table 4.1 for cases using the Brooks-Corey model and in Table 4.2 for the other ones. These parameters have been chosen in such a way that water is more likely to be in RT1 than in RT0: indeed, at a fixed pressure, the water saturation is higher in RT1 than in RT0. This can be observed on the plots of the capillary-pressure functions depicted in Figures 4.12–4.14. On these figures, the relative permeability functions are also shown. Let us, in particular, remark the non-Lipschitz character of the relative permeability in the van Genuchten-Mualem case. Thus, in order to avoid infinite values for the derivative of $k_r(s)$ when $s \to 1 - s_{\text{rn}}$, we approximate it for $s \in [s_{\text{lim}}, 1 - s_{\text{rn}}]$ using a second degree polynomial $\widetilde{k}_r(s)$. This polynomial satisfies the conditions

$$k_r(s_{\text{lim}}) = \widetilde{k}_r(s_{\text{lim}}), \qquad \widetilde{k}_r'(s_{\text{lim}}) = k_r'(s_{\text{lim}}), \qquad \widetilde{k}_r(1 - s_{\text{rn}}) = 1,$$

where s_{lim} is chosen so that $s_{\text{eff}} = 0.998$.

	$1 - s_{\text{rn}}$	s_{rw}	p_b [Pa]	n	λ [m^2]	ϕ
RT0	1.0	0.1	$1.4708 \cdot 10^3$	3.0	10^{-11}	0.35
RT1	1.0	0.2	$3.4301 \cdot 10^3$	1.5	10^{-13}	0.35

Table 4.1: Parameters used for the Brooks-Corey model.

	$1 - s_{\text{rn}}$	s_{rw}	n	λ [m^2]	ξ [m^{-1}]	ϕ
RT0 (Sand)	1.0	0.0782	2.239	$6.3812 \cdot 10^{-12}$	2.8	0.3658
RT1 (Clay)	1.0	0.2262	1.3954	$1.5461 \cdot 10^{-13}$	1.04	0.4686

Table 4.2: Parameters used for the van Genuchten-Mualem model.

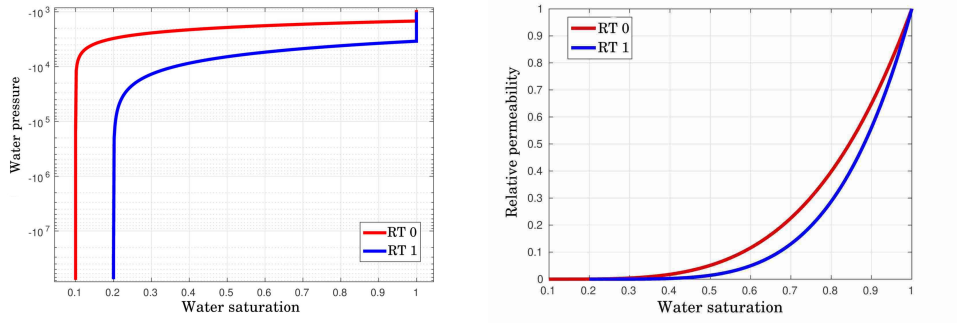


Figure 4.12: Water pressure and relative permeability curves for the Brooks-Corey model.

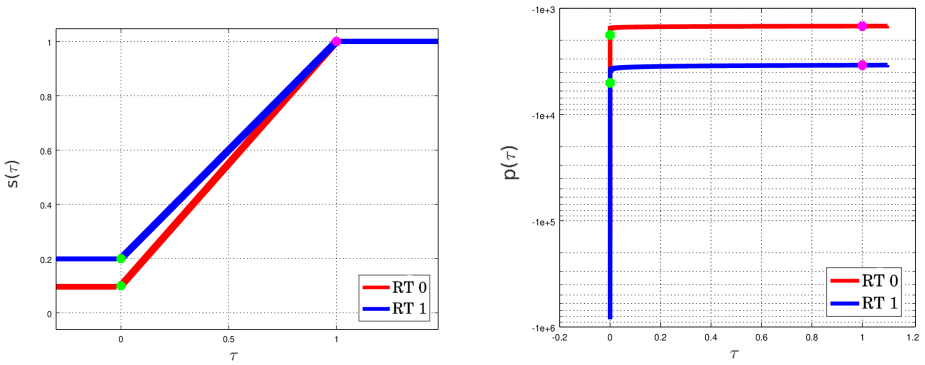


Figure 4.13: Parametrized saturation and pressure functions using the Brooks-Corey model and parameters of Table 4.1. The green dot indicates the value for $\tau = \tau_*$ and the magenta one $\tau = \tau_s$.

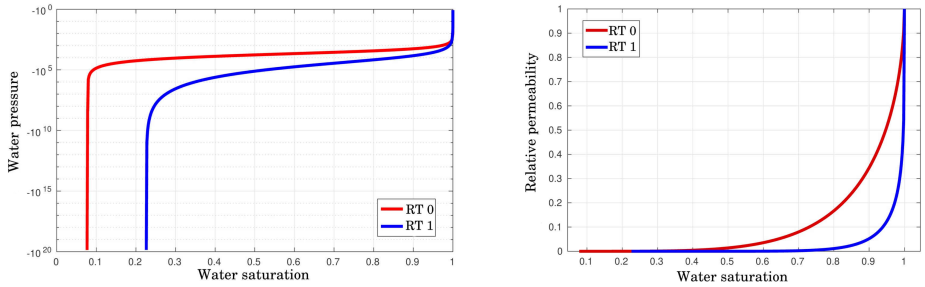


Figure 4.14: Water pressure and relative permeability curves for the van Genuchten-Mualem model.

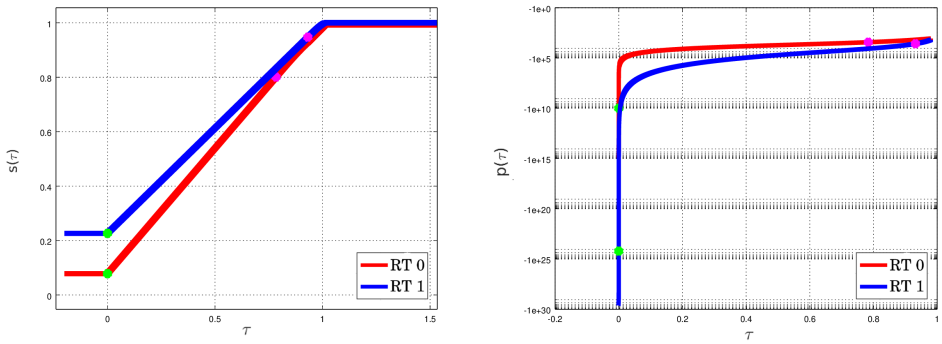


Figure 4.15: Parametrized saturation and pressure functions using the van Genuchten Mualem model using parameters of Table 4.2. The green dot indicate the value for $\tau = \tau_*$ and the magenta one $\tau = \tau_s$.

4.4.2 Comparison of the results in non-steep cases

We now analyze the results obtained on the test cases which were previously introduced. We use uniform time discretizations. The time-step Δt depends on the test case and is reported in Table 4.5 together with the others numerical parameters used for these simulations. The detailed results for the different cases, methods and meshes along with figures of the solutions are reported in Appendix §4.5.

In Table 4.3 we present a brief classification of the proposed methods based on both robustness (R) and accuracy (A) criteria. For each criterion, the colour choice corresponds to the following glossary: green for good, orange for average, red for bad. Regarding the robustness, a non-convergent method is classified as red; orange is used if a method faces many times difficulties during Newton's resolution (maximal number of iterations reached, a much larger number of total iterations in comparison to other methods...); the green label is used in other cases. Thus, a method, having a relative error in the same order as the best performing one, is tagged as green; if an error has one (resp. several) order(s) of magnitude more than the best performing one, the label of the corresponding method is taken as orange (resp. red).

Let us discuss each test case in details, starting with the filling test case simulated with the Brooks-Corey model. Table 4.6 shows that Method A has the smallest saturation relative error with the coarsest mesh. Subsequent refinements then enable to reduce the error related to methods B, C, D at a higher convergence rate, leading to errors on the finest grid that are smaller than the one obtained with the classical scheme A. All the methods face difficulties in the Brooks-Corey filling case for the third mesh and the first time-step. This is due to our non-optimal choice of a uniform time discretization. A simple time-step adaptation strategy similar to the one used in the steep case would fix this issue.

Keeping the Brooks-Corey model, if we now analyze the results reported in Table 4.7 for the drainage case, we notice that methods B, C and D always have a smaller error than method A that converges again at a slower rate. Concerning Method C, it behaves as Method A in terms of accuracy and is fairly cheaper in terms of iterations with respect to this one.

We now consider the results obtained with the van Genuchten-Mualem model. Table 4.8 summarizes the results obtained with the filling case. We can notice that all methods have approximately the same errors and convergence rates. Regarding Newton's cost, the conclusions are similar to the ones made for the Brooks-Corey tests.

	Method A		Method B		Method C		Method D_1		Method D_2	
	R	*A*	*R*	*A*	*R*	*A*	*R*	*A*	*R*	*A*
Brooks-Corey Filling case										
Brooks-Corey Drainage case										
van Genuchten-Mualem Filling case										
van Genuchten-Mualem Drainage case										

Table 4.3: Summary of methods' robustness (*R*) and accuracy (*A*) classification. Method D_1 and Method D_2 denote Method D with Schur complement and bisection method respectively. Color legend: green= good, orange= passing, red=bad.

In the drainage case (see Table 4.9), methods B and C turn out to be more precise and to converge faster that Method A. On the other hand, Methods B and C require more Newton iterations than Method A and D that almost have the same iterations' cost. Moreover we observe that all methods require an important maximum number of iterations to converge which is greater than 50 (reaching the number of 100 iteration for the finer meshes.)

Throughout all these non-steep tests we can also notice that the number of Newton iterations to reach convergence with method B is larger than for the other ones.

Let us now make one last comment on the results obtained using the two proposed double parametrizations (see Eq. (4.3.13)–(4.3.14)) in Method D: they provide the same solutions with the same accuracy in all tests. They only differ in terms of Newton iterations which slightly vary from one parametrization to the other one according to the test case.

4.4.3 Tests with Brooks-Corey model and steep capillary-pressure curves

The aim of this section is to evaluate the robustness of Newton's algorithm when used with the four previous methods and steep capillary-pressure curves. We use the same filling and drainage cases with Brooks-Corey model as in the previous section. We here only change the value of the parameter n which is now equal to 120 for rock-type RT0 and 60 for RT1, making the problem (4.1.1)–(4.1.8) close to a strongly-degenerate parabolic case. The corresponding capillary pressure curves are represented in Figure 4.16. The time evolution for these tests is adaptive:

- **Filling case**
  Minimal, maximal and initial time-steps are such that $\Delta t_{\min} = 10^{-6}s$, $\Delta t_{\max} = 10^{4}s$, $\Delta t^0 = 10^{-6}s$ and, for $n \geqslant 0$,

  $$\Delta t^{n+1} = \min(\Delta t_{\max}, 1.2\Delta t^n)$$

  in case of Newton's convergence or

  $$\Delta t^{n+1} = \max(\Delta t_{\min}, \Delta t^n/2)$$

  in the absence of convergence. In the latter case, for $\Delta t = \Delta t_{\min}$, the simulation stops. $N_{\max} = 30$ is taken as maximal number of Newton's iterations.

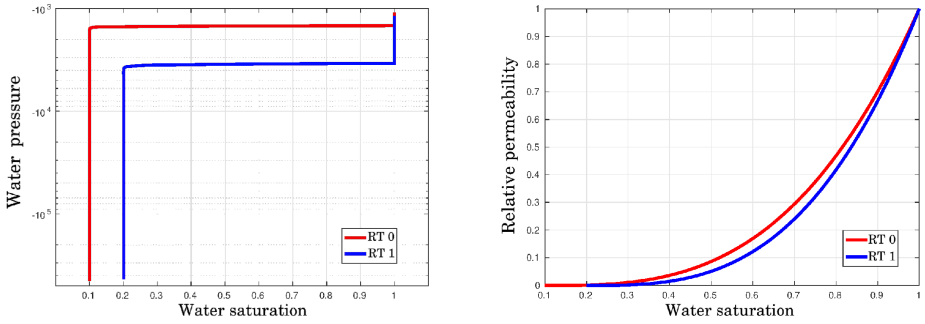


Figure 4.16: Steep cases: water pressure and relative permeability curves for the Brooks-Corey model.

- **Drainage case**
  Minimal, maximal and initial time-steps are such that $\Delta t_{\min} = 1s$, $\Delta t_{\max} = 10^5 s$, $\Delta t^0 = 1s$ and, for $n \geqslant 0$,

$$\Delta t^{n+1} = \begin{cases} 2.5\Delta t^n & \text{if } \Delta t < 500s, \\ \min(\Delta t_{\max}, 1.2\Delta t^n) & \text{otherwise,} \end{cases}$$

  in case of a successful time-step, or

$$\Delta t^{n+1} = \begin{cases} \max(\Delta t_{\min}, \Delta t^n/5) & \text{if } \Delta t < 500s, \\ \Delta t^n/2 & \text{otherwise,} \end{cases}$$

  in the absence of convergence with $N_{\max} = 30$ iterations. If Newton does not converge for $\Delta t = \Delta t_{\min}$ the simulation stops.

4.4.3.1 Comparison of the results

Table 4.4 shows that only methods A, B, C^2, D_1^2 and D_2 converge for all test cases. Here and hereafter, the exponents 1 or 2 refer to the choice of the double parametrization presented in §4.3.3. Figure 4.23 reports on the evolution of the cumulated number of Newton iterations for the filling case with the 50×30 cells mesh. Apart from method D which faces difficulties at the beginning, all curves evolve in the same way. These conclusions remain valid for the 400×240 cells mesh with an exception for method B whose number of iterations increased as it can be observed in Figure 4.24.

Figures 4.26 and 4.27 show the results obtained on the drainage case with meshes of resolutions 50×30 and 400×240 respectively. In both cases, the methods C^1 and D_1^1 face more difficulties to converge than the other ones around time $t = 348500s$. It corresponds to the moment at which the cells line in Ω_2 below the interface between Ω_3 and Ω_2 starts to empty. Note that the number of Newton iterations also increases for method B at that particular time on the finer mesh too. Method C^2 also encounters difficulties on the coarser mesh but at a earlier time (when the cells line in Ω_1 below the interface between Ω_1 and Ω_3 starts to empty) and to a lesser extent. On the whole, the results on this last case show a higher degree of robustness for methods A and the second proposition of the double parametrization which has been designed with the aim of controlling and bounding its derivatives, as it can be seen in Figures 4.17.

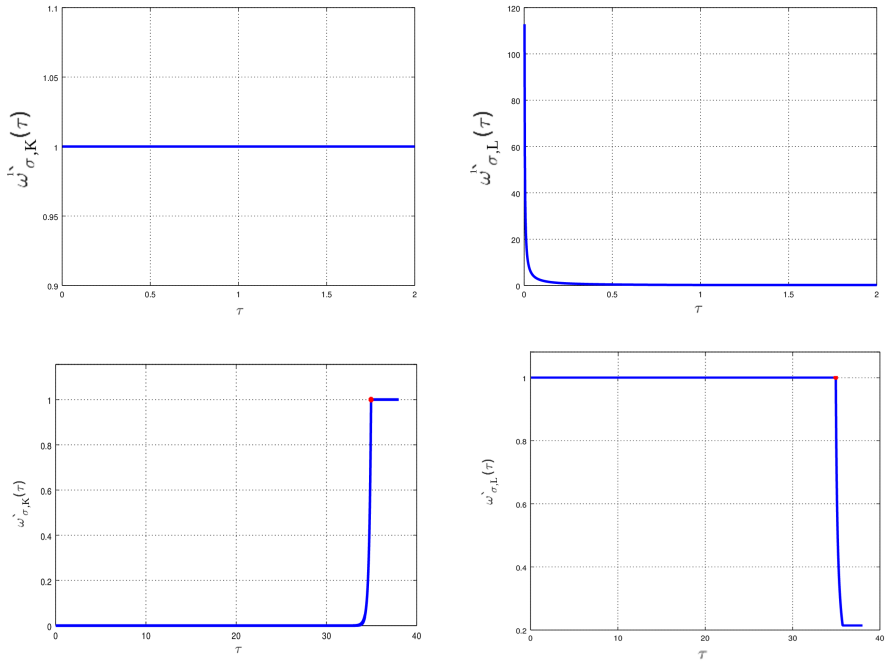


Figure 4.17: Comparison between $\omega'^1_{\sigma,K}(\cdot), \omega'^1_{\sigma,L}(\cdot)$ (above) and $\omega'^2_{\sigma,K}(\cdot), \omega'^2_{\sigma,L}(\cdot)$ (below) when using steep capillary pressure curves.

		Method A	Method B	Method C		Method D_1		Method D_2	
	Mesh			dp_1	dp_2	dp_1	dp_2	dp_1	dp_2
Filling case	50×30								
	400×240								
Drainage case	50×30								
	400×240								

Table 4.4: Summary of methods' robustness classification for steep tests. Method D_1 and Method D_2 denote Method D with Schur complement and bisection method respectively. Color legend: green= good, orange= passing, red=not converge.

4.4.4 Overall method evaluation

Using this glossary and the results obtained in the steep and non-steep cases, we proceed, in the following of this section, to a general evaluation of the five studied methods.

Let us start with Method A. In this approach, rock-type interface faces are treated like classical inner faces and the pressure continuity on these interfaces is not enforced. Nevertheless, if the simulation is performed on a sufficiently refined mesh, a good approximation of this condition can be obtained. In the previous tests, and in particular in the steep ones (see §4.4.3), this method turns out to be very robust. On fine meshes its accuracy is, in general, close to the ones of other methods. In the filling case with the Brooks-Corey model (see Table 4.6), this method is even the most accurate one on coarse meshes. A noticeable drawback of this method is the loss of the linear convergence rate when used with the Brooks-Corey model (see Tables 4.6–4.7).

Method B is the first approach we propose with a specific treatment for the rock-type interfaces, which only entails moderate changes in terms of implementation compared to method A. We here just add two thin cells around the rock-type interfaces and neglect, for these new cells, the fluxes through the faces with small measures. It features a rather good robustness since it also converges in the steep cases (see §4.4.3). For non-steep cases, it always provides a rather accurate solution and good robustness. Compared to method A, the linear convergence rate is recovered at the price of about 10% extra Newton iterations.

Method C is the first method which strongly enforces the pressure continuity on the rock-type interfaces. Here, the interfaces are thickened in thin cells and the pressure continuity is ensured by introducing a second parametrization (4.3.12) for which we propose two different forms detailed in Equations (4.3.13)–(4.3.14). This parametrization should be calculated beforehand and depends on the chosen petro-physical model. Thus, this method involves more changes for its implementation with respect to the previous one. In non-steep simulations, the two proposed double parametrizations provide the same solutions with the same accuracy with just a slight difference in the required number of iterations. Moreover it behaves as Method B in all non-steep simulations in terms of error. In the steep tests a remarkable difference of performance between the use of the two proposed parametrizations for the pressure continuity at interfaces arises: the first proposition of parametrization converges only in drainage case while the second one always converges showing, generally, a competitive robustness.

The last studied method Method D guarantees the flux conservation between all cells of the initial meshes and pressure continuity at rock-type interfaces. As for Method C, it also uses a double parametrization (4.3.12). In the non-steep cases, the application of the two proposed second parametrizations, as in Method C, provides the same solutions with the same accuracy with just a slight difference in the required number of iterations, as already remarked in §4.4.2. Moreover, Methods D_1 and D_2 show in all tests a good robustness and the same rather good accuracy: in drainage cases their accuracy is fairly better than Method A when using the Brooks-Corey model and, when employing the van Genuchten-Mualem model, they show a relative error almost halved with respect to the one of Method A. In filling cases they have almost the same accuracy as Method A. Methods D_1 and D_2 always recover a first-order convergence except for the drainage test case with the van Genuchten-Mualem model in which the convergence rate is slightly degraded. The fact that Methods D_1 and D_2 show the same accuracy is not surprising: the only difference between the two methods is how we solve the system that, for its part, does not change. In steep tests both double parametrizations employed in Method D_1 and D_2 make the simulation converge, apart for the filling test case in which Method D_1^1 fails, just showing in some cases a difference of behaviour in terms of robustness as detailed in §4.4.3.1.

Finally we can conclude that if we want to perform simulation for test cases with steep pressure curves, we can choose between Method A or B or, if one does not mind making larger code changes, methods C^2, D_1^2 or D_2 can also be used. If it is not the case, for coarse meshes, Method A ensures a good robustness and an accurate solution without any particular treatment for interfaces. For more refined meshes, even if its accuracy is slightly lower or comparable -it depends on the specific test case- to that of Method D, Method B is easier to implement and the least intrusive with respect to methods introducing a treatment for interfaces. So, if the choice is based on an accuracy criterion actually Method B, C and D are almost equivalent; in terms of ease of implementation the best choice is Method B.

4.5 Figures and data related to the non-steep cases

	Δt	τ_*	ϵ	ϵ_{bis}	γ	δ_B	δ_C
Filling – Brooks-Corey	500	10^{-10}	10^{-12}	10^{-16}	10^{-15}	10^{-6}	$2 \cdot 10^{-6}$
Filling – van Genuchten-Mualem	500	10^{-8}	10^{-12}	10^{-16}	10^{-15}	10^{-6}	$2 \cdot 10^{-6}$
Drainage – Brooks-Corey	1000	10^{-10}	10^{-12}	10^{-16}	10^{-15}	10^{-6}	$2 \cdot 10^{-6}$
Drainage – van Genuchten-Mualem	1000	10^{-8}	10^{-12}	10^{-16}	10^{-15}	10^{-6}	$2 \cdot 10^{-6}$

Table 4.5: Numerical parameters used in the examples.

4.5.1 Filling case using Brooks-Corey model

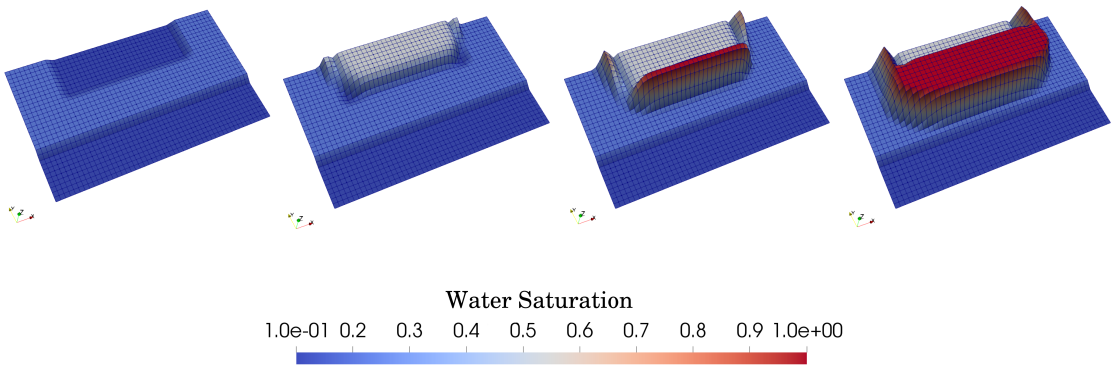


Figure 4.18: Evolution of the saturation profile for $t \in \{0\,\text{s}, 21.5 \cdot 10^3\,\text{s}, 41.5 \cdot 10^3\,\text{s}, 86.4 \cdot 10^3\,\text{s}\}$ for the non-steep filling case, using Brooks-Corey model, Method B and the 50×30 cells mesh.

Method A	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	9.60719e-2	8.08028e-2	6.41616e-2	5.18869e-2
Rate of convergence	–	0.25	0.333	0.306
# total iterations	647	777	1074	1236
# avg iterations	3	4	6	7
# max iterations	18	21	168	32
Method B	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.43731e-1	1.0421e-1	6.3325e-2	2.76736e-2
Rate of convergence	–	0.464	0.719	1.194
# total iterations	835	959	1279	1428
# avg iterations	4	5	7	8
# max iterations	19	21	168	38
Method C	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.46706e-1	1.06227e-1	6.45985e-2	2.84733e-2
Rate of convergence	–	0.465	0.7156	1.182
# total iterations	690	796(794)	1106(1102)	1253(1247)
# avg iterations	3	4	6	7
# max iterations	20(18)	21	168	29
Method D_1	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.7701e-1	1.26129e-1	7.75469e-2	3.66345e-2
Rate of convergence	–	0.489	0.702	1.082
# total iterations	620(634)	721(734)	1001(1018)	1140(1146)
# avg iterations	3	4	5	6
# max iterations	17	20	155	32
Method D_2	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.7701e-1	1.26129e-1	7.75469e-2	3.66345e-2
Rate of convergence	–	0.489	0.702	1.082
# total iterations	590	714	999	1140
# avg iterations	3	4	5	6
# max iterations	17	20	155	32
# avg it. bisection per face	16	15	15	14

Table 4.6: Results for the non-steep filling case using Brooks-Corey model. For methods C and D_1, we specify within parentheses the number of iterations corresponding to the second choice of double parametrization when it differs from the one obtained with the first choice, which is reported without parentheses.

4.5.2 Drainage case using Brooks-Corey model

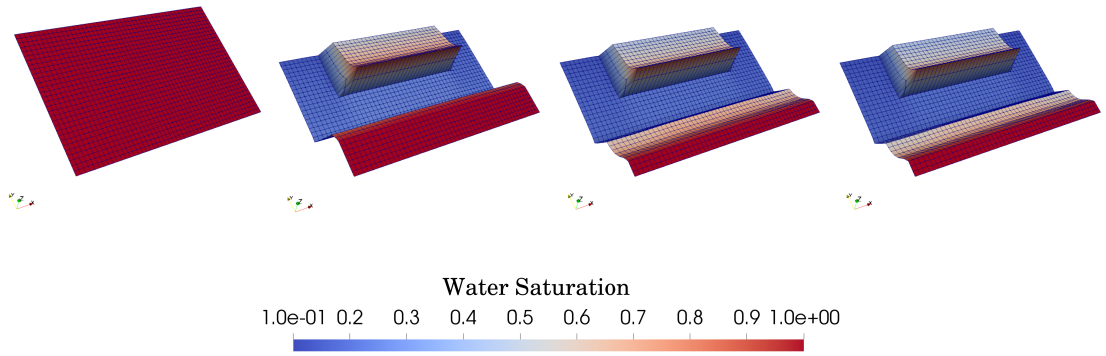


Figure 4.19: Evolution of the saturation profile for $t \in \{0\,\mathrm{s}, 35 \cdot 10^4\,\mathrm{s}, 70 \cdot 10^4\,\mathrm{s}, 105 \cdot 10^4\,\mathrm{s}\}$ for the non-steep drainage case, using Brooks-Corey model, Method B and the 50×30 cells mesh.

Method A	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	4.48867e-2	2.60531e-2	1.64213e-2	1.110698e-2
Rate of convergence	–	0.785	0.666	0.564
# total iterations	2598	2848	3258	3819
# avg iterations	2	2	3	3
# max iterations	21	24	29	32
Method B	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.80469e-2	9.9613e-3	4.83626e-3	1.77811e-3
Rate of convergence	–	0.857	1.042	1.443
# total iterations	2845	3056	3448	3918
# avg iterations	2	2	3	3
# max iterations	20	24	20	32
Method C	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.81634e-2	1.00638e-2	4.92295e-3	1.84945e-3
Rate of convergence	–	0.851	1.032	1.412
# total iterations	2659(2653)	2893(2887)	3304(3298)	3804(3798)
# avg iterations	2	2	3	3
# max iterations	20(21)	24	28(29)	32
Method D_1	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	3.03634e-2	1.62917e-2	8.64114e-3	4.18359e-3
Rate of convergence	–	0.898	0.915	1.046
# total iterations	2659(2665)	2919(2905)	3329(3324)	3863(3861)
# avg iterations	2	2	3	3
# max iterations	21(20)	24	29(30)	32
Method D_2	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	3.03634e-2	1.62917e-2	8.64114e-3	4.18359e-3
Rate of convergence	–	0.898	0.915	1.046
# total iterations	2614	2894	3320	3856
# avg iterations	2	2	3	3
# max iterations	21	24	29	32
# avg it. bisection per face	35	34	33	32

Table 4.7: Results for the non-steep drainage case using Brooks-Corey model. For methods C and D_1, we specify within parentheses the number of iterations corresponding to the second choice of double parametrization when it differs from the one obtained with the first choice, which is reported without parentheses.

4.5.3 Filling case using van Genuchten-Mualem model

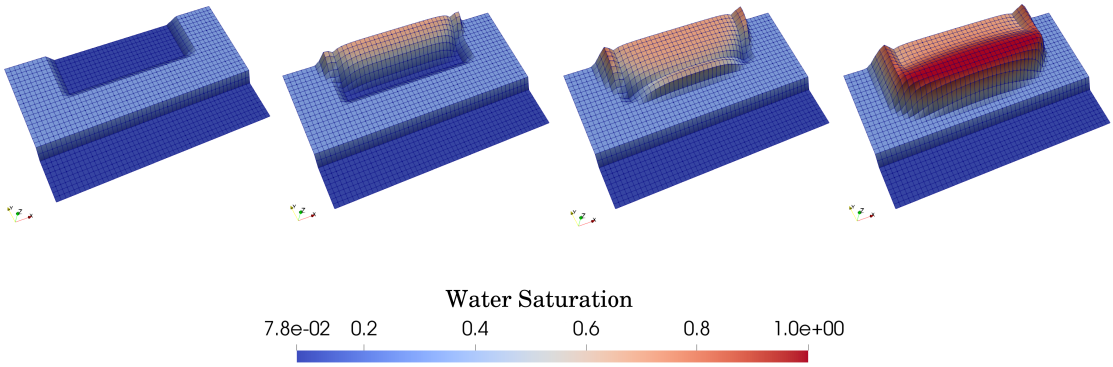


Figure 4.20: Evolution of the saturation profile for $t \in \{0\,\mathrm{s}, 21.5 \cdot 10^3\,\mathrm{s}, 41.5 \cdot 10^3\,\mathrm{s}, 86.4 \cdot 10^3\,\mathrm{s}\}$ for the non-steep filling case using van Genuchten-Mualem model, Method B and the 50×30 cells mesh.

Method A	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.05534e-1	7.48124e-2	4.55216e-2	2.1125e-2
Rate of convergence	–	0.496	0.717	1.108
# total iterations	575	667	782	930
# avg iterations	3	3	4	5
# max iterations	9	12	15	18
Method B	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.23187e-1	8.67715e-2	5.30592e-2	2.40712e-2
Rate of convergence	–	0.506	0.71	1.14
# total iterations	836	900	959	1076
# avg iterations	4	5	5	6
# max iterations	9	12	15	18
Method C	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.23321e-1	8.68681e-2	5.31222e-2	2.41106e-2
Rate of convergence	–	0.509	0.706	1.14
# total iterations	571(573)	678(675)	779(800)	934(960)
# avg iterations	3	3	4	5
# max iterations	9	12	15	18
Method D_1	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.50856e-1	1.0295e-1	6.20923e-2	2.91945e-2
Rate of convergence	–	0.551	0.729	1.089
# total iterations	579(581)	677(676)	785(814)	933(985)
# avg iterations	3	3	4	5
# max iterations	9	12	15	18
Method D_2	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.50856e-1	1.0295e-1	6.20923e-2	2.91945e-2
Rate of convergence	–	0.551	0.729	1.089
# total iterations	579	674	783	933
# avg iterations	3	3	4	5
# max iterations	9	12	15	18
# avg it. bisection per face	15	13	12	11

Table 4.8: Results for the non-steep filling case using van Genuchten-Mualem model. For methods C and D_1, we specify within parentheses the number of iterations corresponding to the second choice of double parametrization when it differs from the one obtained with the first choice, which is reported without parentheses.

4.5.4 Drainage case using van Genuchten-Mualem model

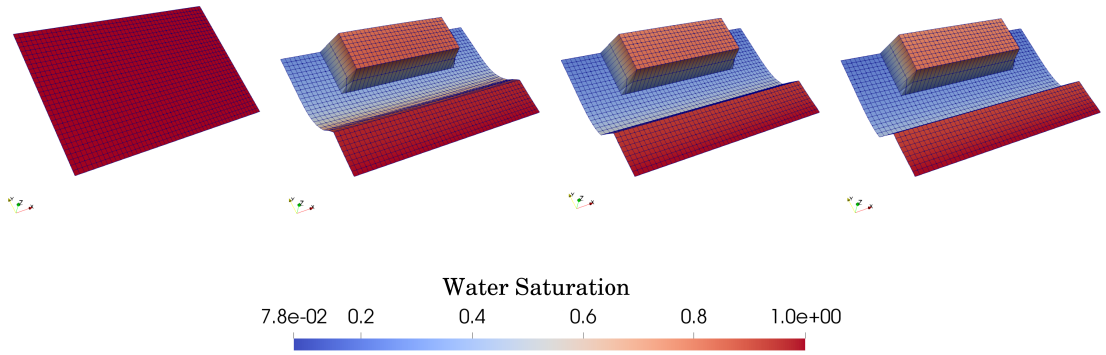


Figure 4.21: Evolution of the saturation profile for $t \in \{0s, 35 \cdot 10^4 s, 70 \cdot 10^4 s, 105 \cdot 10^4 s\}$ for the non-steep drainage case using the van Genuchten-Mualem model, Method B and the 50×30 cells mesh.

Method A	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	1.50494e-2	7.3434e-3	3.57016e-3	1.66693e-3
Rate of convergence	–	1.035	1.04	1.099
# total iterations	2333	2330	2325	2326
# avg iterations	2	2	2	2
# max iterations	65	67	67	72
Method B	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	6.52099e-3	3.11282e-3	1.35196e-3	4.53001e-4
Rate of convergence	–	1.064	1.203	1.577
# total iterations	2949	3006	3028	3236
# avg iterations	2	2	2	3
# max iterations	97	96	96	100
Method C	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	6.52104e-3	3.11287e-3	1.35201e-3	4.53048e-4
Rate of convergence	–	1.067	1.203	1.577
# total iterations	2855	2818(2817)	2904(2902)	2970(2962)
# avg iterations	2	2	2	2
# max iterations	99	98(97)	97(95)	101(93)
Method D_1	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	6.58861e-3	3.56638e-3	1.87745e-3	1.01069e-3
Rate of convergence	–	0.886	0.926	0.893
# total iterations	2349	2351(2350)	2350	2367
# avg iterations	2	2	2	2
# max iterations	76	77(76)	80	84
Method D_2	50×30	100×60	200×120	400×240
$\frac{\|s-s_{\text{ref}}\|_{L^2([0,T],\Omega)}}{\|s_{\text{ref}}\|_{L^2([0,T],\Omega)}}$	6.58861e-3	3.56638e-3	1.87745e-3	1.01069e-3
Rate of convergence	–	0.886	0.923	0.893
# total iterations	2348	2345	2350	2367
# avg iterations	2	2	2	2
# max iterations	76	76	80	84
# avg it. bisection per face	34	33	32	31

Table 4.9: Results for the non-steep drainage case using van Genuchten-Mualem model. For methods C and D_1, we specify within parentheses the number of iterations corresponding to the second choice of double parametrization when it differs from the one obtained with the first choice, which is reported without parentheses.

4.6 Figures and data related to the steep cases

4.6.1 Filling case

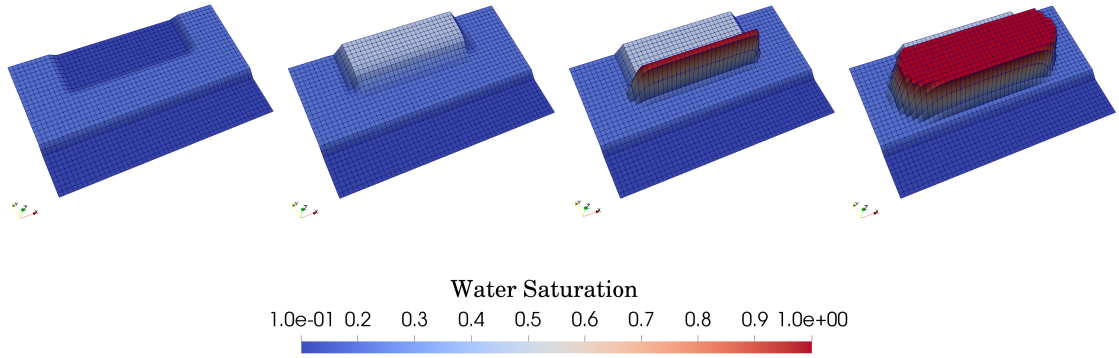


Figure 4.22: Evolution of the saturation profile for $t \in \{0\,\text{s}, 5422.843\,\text{s}, 37844.5\,\text{s}, 86.4 \cdot 10^3\,\text{s}\}$ for the steep filling case using Brooks-Corey model, Method B and the 50×30 cells mesh

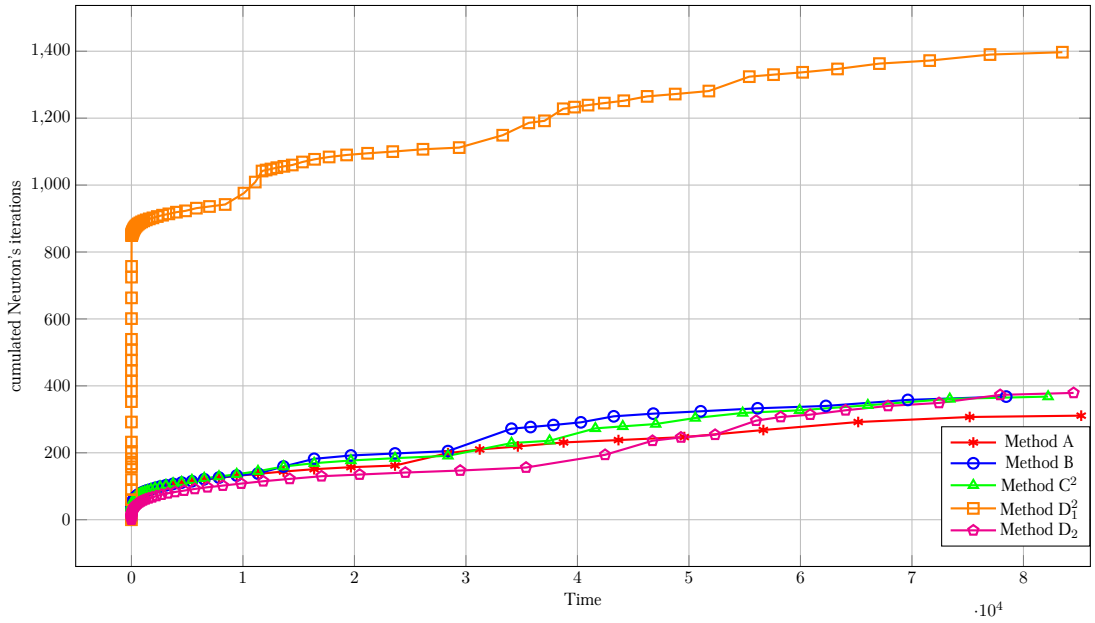


Figure 4.23: Steep filling case: Evolution of the cumulated number of Newton's iterations for the 50×30 cells mesh

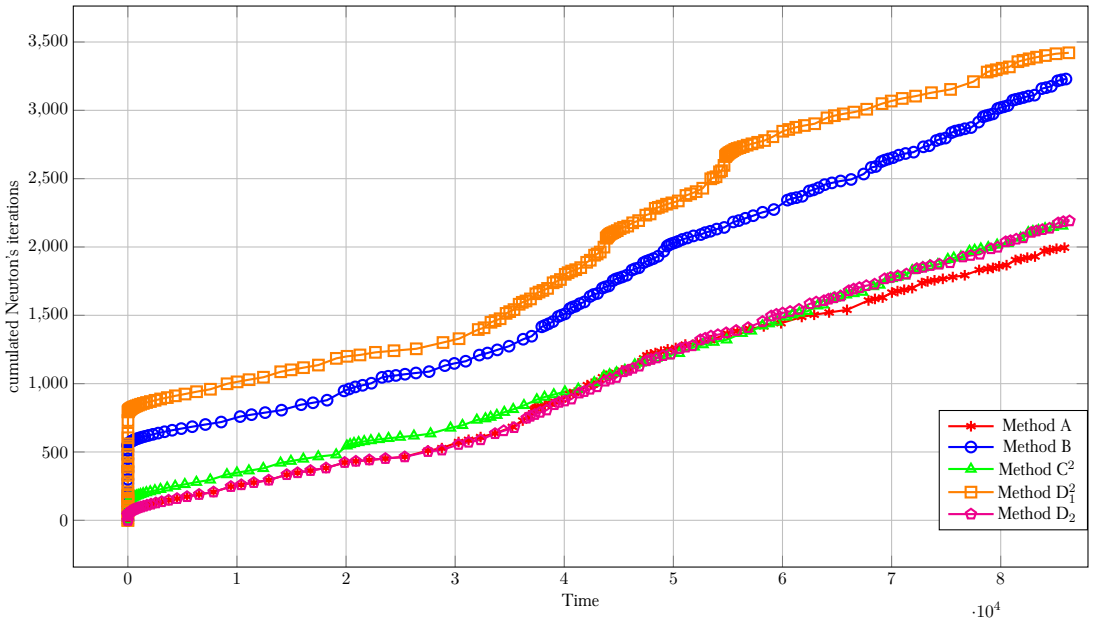


Figure 4.24: Steep filling case: Evolution of the cumulated number of Newton's iterations for the 400×240 cells mesh

4.6.2 Drainage case

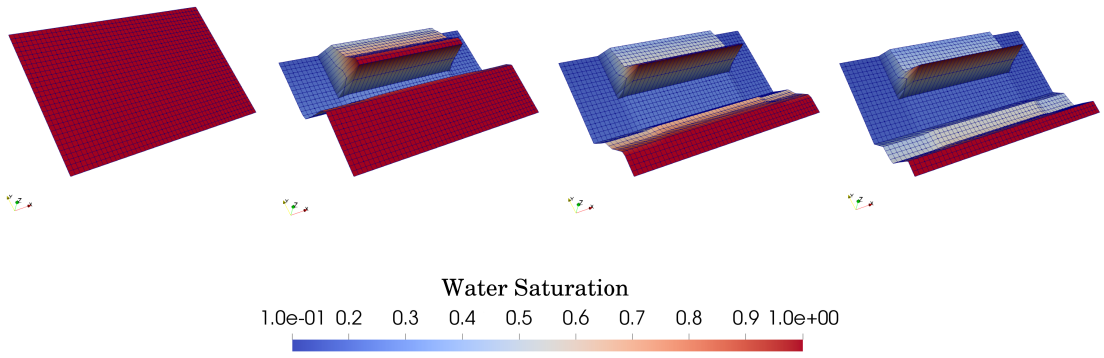


Figure 4.25: Evolution of the saturation profile for $t \in \{0\,\text{s}, 81593.8\,\text{s}, 308776\,\text{s}, 105 \cdot 10^4\,\text{s}\}$ for the steep drainage case using Brooks-Corey model, Method B and the 50×30 cells mesh

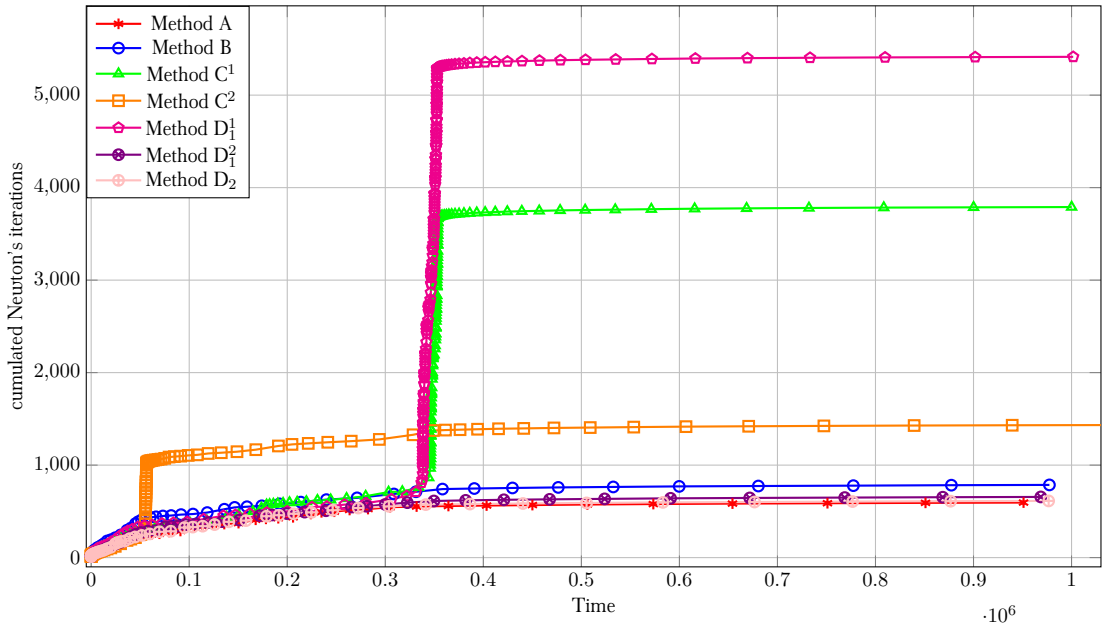


Figure 4.26: Steep drainage case: Evolution of the cumulated number of Newton's iterations for the 50×30 cells mesh

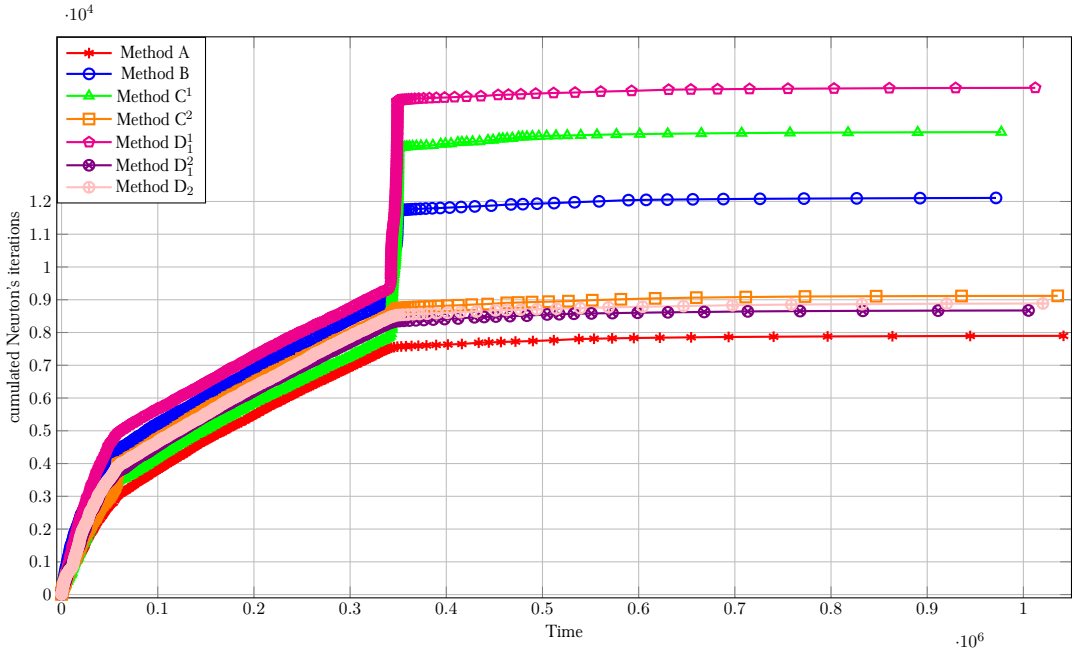


Figure 4.27: Steep drainage case: Evolution of the cumulated number of Newton's iterations for the 400×240 cells mesh

Chapter 5

Finite volume scheme and numerical strategies to solve two-phase Darcy flows in heterogeneous domains

Relying on the ideas developed in the previous chapters for the Richards equation, we are now in a position to cope with the more difficult case of the more difficult model, namely, the immiscible incompressible two-phase system (1.2.1)–(1.2.6) —also known as the *isotherm Dead Oil model* in the reservoir engineering community— in a heterogeneous domain.

After introducing the finite volume discretization (§5.1) and extending the specific techniques addressing stiffness and interface transmission to the two-phase system (§5.2), we perform three CO_2 injection test cases (§5.3) that are representative of realistic operating conditions in order to validate the numerical strategies proposed.

5.1 Finite volume scheme for the two-phase system

5.1.1 State of the art

Similarly to the Richards equation, it is useful to be aware of the vast amount of literature on the numerical resolution of the immiscible incompressible two-phase flow system. The review below is in no way exhaustive.

For a homogeneous capillary pressure function, we can cite the seminal work of Chen and Ewing [53, 54] and the analysis of Radu et al. [119] using finite elements, the contribution of Eymard et al. [76] using finite volumes, or the more recent multinumerics method by Doyle et al. [59].

The case of a heterogeneous medium has received a lot of attention with various methods of the discontinuous Galerkin family [68, 77, 97, 111], the finite element [90, 91, 96] and the finite volume one [30, 65, 74]. In particular, the treatment of discontinuity in capillary pressure has been the subject of extensive researches in the VAG (Vertex Approximate Gradient) community [32, 34, 38]. This occurs in conjunction with a renewed interest in the HU (Hybrid Upwiding) paradigm [6, 88] for matrix-fracture systems.

In the sequel, only the TPFA finite volume scheme will be needed to illustrate how our numerical strategies can be deployed.

5.1.2 Implicit TPFA discretization of the model

Let us first rewrite the system (1.2.1)–(1.2.6) under the slightly more condensed form

$$\partial_t(\phi s_\alpha) + \nabla \cdot v_\alpha = 0 \qquad \text{in } Q_T, \tag{5.1.1a}$$
$$v_\alpha + \lambda \frac{k_{r,\alpha}(s_\alpha, x)}{\mu_\alpha}(\nabla p_\alpha - \varrho_\alpha g) = 0 \qquad \text{in } Q_T, \tag{5.1.1b}$$
$$p_{\rm nw} - p_{\rm w} = p_{\rm c} \qquad \text{in } Q_T, \tag{5.1.1c}$$
$$s_{\rm nw} - \mathcal{S}_{\rm nw}(p_{\rm c}, x) = 0 \qquad \text{in } Q_T, \tag{5.1.1d}$$
$$s_{\rm w} + s_{\rm nw} = 1 \qquad \text{in } Q_T. \tag{5.1.1e}$$
$$p_\alpha = p_\alpha^{\rm D} \qquad \text{on } \Gamma^{\rm D} \times (0,T), \tag{5.1.1f}$$
$$v_\alpha \cdot \nu = q_\alpha^{\rm N} \qquad \text{on } \Gamma^{\rm N} \times (0,T), \tag{5.1.1g}$$
$$s_\alpha(\cdot, t=0) = s_\alpha^0 \qquad \text{in } \Omega, \tag{5.1.1h}$$

where $\alpha \in \{\rm w, nw\}$. Given an admissible mesh and a sequence of time-steps as described in §4.2.1, we follow the standard procedure detailed in §4.2.2. Put another way, we integrate each volume balance equation (5.1.1a) over a cell K and apply Green's theorem to transform the volume integral of $\nabla \cdot v_\alpha$ into a close surface integral involving its normal component $v_\alpha \cdot \nu$.

Being implicitly understood that $s_{\rm w} = 1 - s_{\rm nw}$, the resulting scheme reads

$$m_K \phi_K \frac{(s_{\rm w})_K^n - (s_{\rm w})_K^{n-1}}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} (F_{\rm w})_{K\sigma}^n = 0, \tag{5.1.2a}$$
$$m_K \phi_K \frac{(s_{\rm nw})_K^n - (s_{\rm nw})_K^{n-1}}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_K} (F_{\rm nw})_{K\sigma}^n = 0, \tag{5.1.2b}$$
$$(p_{\rm c})_K^n - [(p_{\rm nw})_K^n - (p_{\rm w})_K^n] = 0, \tag{5.1.2c}$$
$$(s_{\rm nw})_K^n - (\mathcal{S}_{\rm nw})_K((p_{\rm c})_K^n) = 0, \tag{5.1.2d}$$

for each cell $K \in \mathscr{T}$, where the numerical fluxes

$$(F_{\rm w})_{K\sigma}^n = \begin{cases} A_\sigma \dfrac{(k_{r,{\rm w}})_\sigma^n}{\mu_{\rm w}} [(\vartheta_{\rm w})_K^n - (\vartheta_{\rm w})_{K\sigma}^n] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm D}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} {\rm d}t \int_\sigma q_{\rm w}^{\rm N} \, {\rm d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \end{cases} \tag{5.1.3a}$$

$$(F_{\rm nw})_{K\sigma}^n = \begin{cases} A_\sigma \dfrac{(k_{r,{\rm nw}})_\sigma^n}{\mu_{\rm nw}} [(\vartheta_{\rm nw})_K^n - (\vartheta_{\rm nw})_{K\sigma}^n], & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm D}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} {\rm d}t \int_\sigma q_{\rm nw}^{\rm N} \, {\rm d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \end{cases} \tag{5.1.3b}$$

are implicit TPFA approximations of $\int_\sigma v_{\rm w} \cdot \nu_{K,\sigma} \, {\rm d}\gamma$ and $\int_\sigma v_{\rm nw} \cdot \nu_{K,\sigma} \, {\rm d}\gamma$. We recall that

$$A_\sigma = \begin{cases} \dfrac{m_\sigma}{\mu} \dfrac{\lambda_K \lambda_L}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} & \text{if } \ \sigma = K|L \in \mathscr{E}_K \cap \mathscr{E}_{\rm int}, \\ \dfrac{m_\sigma}{\mu} \dfrac{\lambda_K}{d_{K,\sigma}} & \text{if } \ \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}, \end{cases} \tag{5.1.4}$$

is the transmissibilty, as was defined in (2.1.8), while $w_{K\sigma}$ stands for the mirror value of w_K as defined in the previous chapters. In (5.1.3), the face mobilities are upwinded according to the sign

of the difference in the hydraulic heads

$$(\vartheta_{\rm w})^n = (p_{\rm w})^n \ + \psi_{\rm w} \ = (p_{\rm w})^n \ - \varrho_{\rm w}\, g\cdot x, \tag{5.1.5a}$$
$$(\vartheta_{\rm nw})^n = (p_{\rm nw})^n + \psi_{\rm nw} = (p_{\rm nw})^n - \varrho_{\rm nw}\, g\cdot x. \tag{5.1.5b}$$

In other words,

$$(k_{r,{\rm w}})_\sigma^n = \begin{cases} (k_{r,{\rm w}})_K((s_{\rm w})_K^n) & \text{if } (\vartheta_{\rm w})_K^n - (\vartheta_{\rm w})_{K\sigma}^n \geqslant 0, \\ (k_{r,{\rm w}})_{K\sigma}((s_{\rm w})_{K\sigma}^n) & \text{if } (\vartheta_{\rm w})_K^n - (\vartheta_{\rm w})_{K\sigma}^n < 0. \end{cases} \tag{5.1.6a}$$
$$(k_{r,{\rm nw}})_\sigma^n = \begin{cases} (k_{r,{\rm nw}})_K((s_{\rm w})_K^n) & \text{if } (\vartheta_{\rm nw})_K^n - (\vartheta_{\rm nw})_{K\sigma}^n \geqslant 0, \\ (k_{r,{\rm nw}})_{K\sigma}((s_{\rm w})_{K\sigma}^n) & \text{if } (\vartheta_{\rm nw})_K^n - (\vartheta_{\rm nw})_{K\sigma}^n < 0. \end{cases} \tag{5.1.6b}$$

5.2 Application of previously developed techniques

5.2.1 Parametrization by variable switching

Let us recall the hydraulic model for the two-phase flow (1.2.21)–(1.2.22) presented in §1. Setting the effective saturation as

$$\widetilde{s}_{\rm eff} := \widetilde{s}_{\rm eff}(s_{\rm nw}) = \Pi_{[0,1]}\left(\frac{(1-s_{\rm rw})-s_{\rm nw}}{(1-s_{\rm rw})-s_{\rm rn}}\right) = \Pi_{[0,1]}\left(\frac{s_{\rm w}-s_{\rm rw}}{(1-s_{\rm rn})-s_{\rm rw}}\right), \tag{5.2.1}$$

where $\Pi_{[0,1]}$ stands for the projection on $[0,1]$, we have

- for the Brooks-Corey model:

$$k_{r,{\rm w}}(s_{\rm nw}) = \widetilde{s}_{\rm eff}^{\,3+2/n}, \tag{5.2.2a}$$
$$k_{r,{\rm nw}}(s_{\rm nw}) = (1-\widetilde{s}_{\rm eff})^2(1-\widetilde{s}_{\rm eff}^{\,1+2/n}), \tag{5.2.2b}$$
$$\mathcal{S}_{\rm nw}(p_{\rm c}) = \begin{cases} 1-\left[s_{\rm rw}+(1-s_{\rm rn}-s_{\rm rw})\left(\dfrac{p_{\rm c}}{p_b}\right)^{-n}\right] & \text{if } p_{\rm c} > p_b, \\ s_{\rm rn} & \text{if } p_{\rm c} \leqslant p_b; \end{cases} \tag{5.2.2c}$$

- for the van Genuchten-Mualem model:

$$k_{r,{\rm w}}(s_{\rm nw}) = \widetilde{s}_{\rm eff}^{\,1/2}\{1-\big[1-\widetilde{s}_{\rm eff}^{\,1/m}\big]^m\}^2, \tag{5.2.3a}$$
$$k_{r,{\rm nw}}(s_{\rm nw}) = (1-\widetilde{s}_{\rm eff})^{1/2}\big[1-\widetilde{s}_{\rm eff}^{\,1/m}\big]^{2m}, \tag{5.2.3b}$$
$$\mathcal{S}_{\rm nw}(p_{\rm c}) = \begin{cases} 1-\left[s_{\rm rw}+(1-s_{\rm rn}-s_{\rm rw})\left(1+\left|\dfrac{\xi p_{\rm c}}{\varrho_{\rm w} g}\right|^n\right)^{-m}\right] & \text{if } p_{\rm c} > 0, \\ s_{\rm rn} & \text{if } p_{\rm c} \leqslant 0, \end{cases} \tag{5.2.3c}$$

  with $m = 1-1/n$.

As we did in §4.4.1.3, in order to avoid infinite values for the derivative of $k_{r,{\rm w}}(s_{\rm nw})$ when $s_{\rm nw} \to s_{\rm rn}$, we approximate it for $s \in [s_{\rm lim}, 1-s_{\rm rn}]$ using a second degree polynomial $\widetilde{k}_{r,{\rm w}}(\cdot)$.

	$1 - s_{\mathrm{rn}}$	s_{rw}	p_b [Pa]	n
RT0	1.0	0.1	$1.4708 \cdot 10^3$	3.0
RT1	1.0	0.2	$3.4301 \cdot 10^3$	1.5

Table 5.1: Parameters used for the Brooks-Corey model.

	$1 - s_{\mathrm{rn}}$	s_{rw}	n	ξ [m^{-1}]
RT0 (Sand)	1.0	0.0782	2.239	2.8
RT1 (Clay)	1.0	0.2262	1.3954	1.04

Table 5.2: Parameters used for the van Genuchten-Mualem model.

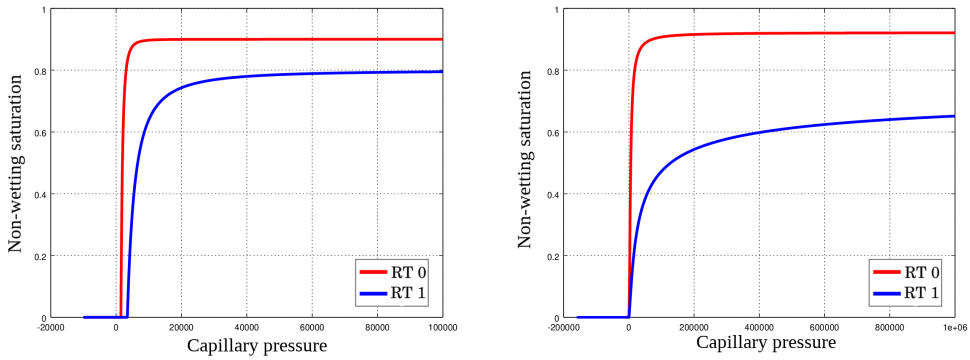


Figure 5.1: Plot of $\mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}})$ curve for the Brooks-Corey model (left) and for the van Genuchten-Mualem model (right).

This polynomial satisfies the conditions

$$k_{r,\mathrm{w}}(s_{\mathrm{nw,lim}}) = \widetilde{k}_{r,\mathrm{w}}(s_{\mathrm{nw,lim}}), \tag{5.2.4a}$$

$$\widetilde{k}'_{r,\mathrm{w}}(s_{\mathrm{nw,lim}}) = k'_{r,\mathrm{w}}(s_{\mathrm{nw,lim}}), \tag{5.2.4b}$$

$$\widetilde{k}_{r,\mathrm{w}}(s_{\mathrm{rn}}) = 1, \tag{5.2.4c}$$

where $s_{\mathrm{nw,lim}}$ is chosen so that $\widetilde{s}_{\mathrm{eff}} = 0.002$.

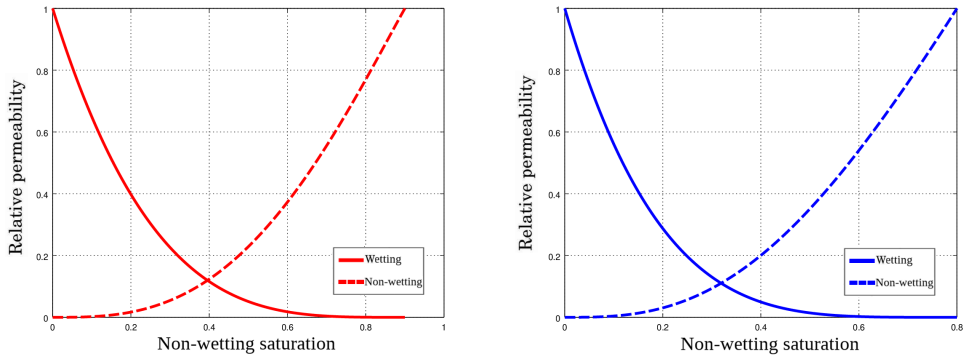
Let us now adapt the parametrization presented in our article [19] to the two-phase flow case. Removing the subscript i related to the rock-type for convenience, the idea is to choose a parametrization of the graph $\{p_{\mathrm{c}}, \mathcal{S}(p_{\mathrm{c}})\}$, i.e., to construct two monotone functions

$$\mathfrak{s}_{\mathrm{nw}} : I \to [s_{\mathrm{rn}}, 1 - s_{\mathrm{rw}}], \qquad \mathfrak{p}_{\mathrm{c}} : I \to \mathbb{R},$$

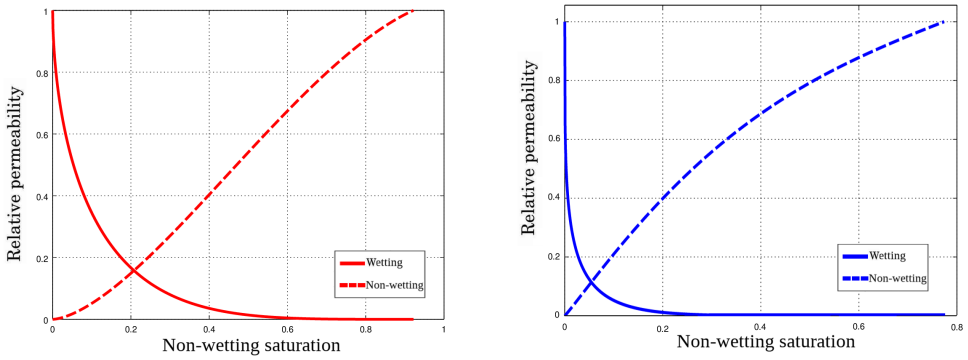
such that

$$\mathfrak{s}_{\mathrm{nw}}(\tau) = \mathcal{S}_{\mathrm{nw}}(\mathfrak{p}_{\mathrm{c}}(\tau)), \qquad 0 < \mathfrak{s}'_{\mathrm{nw}}(\tau) + \mathfrak{p}'_{\mathrm{c}}(\tau) < \infty, \tag{5.2.5}$$

for all $\tau \in I \subset \mathbb{R}$. The latter non-degeneracy assumption ensures that for all $p \in \mathbb{R}$, there exists a unique $\tau \in \mathbb{R}$ such that $(p_{\mathrm{c}}, \mathcal{S}_{\mathrm{nw}}(p_{\mathrm{c}})) = (\mathfrak{p}_{\mathrm{c}}(\tau), \mathfrak{s}_{\mathrm{nw}}(\tau))$. With respect to the parametrization proposed in [19], basically we have to reverse the order of the switch points to follow the behaviour



(a) Plots using the Brooks-Corey model, for rock type 0 (left) and rock type 1 on (right).



(b) Plots using the van Genuchten-Mualem model, for rock type 0 (left) and rock type 1 (right).

Figure 5.2: Plot of the wetting and non-wetting relative permeability curve using the Brooks-Corey model and the van Genuchten-Mualem model.

of law $\mathcal{S}_{\rm nw}$. We set $I = \mathbb{R}$ and

$$\mathfrak{s}_{\rm nw}(\tau) = \begin{cases} \mathcal{S}_{\rm nw}(p_{\rm s} + \varsigma(\tau - \tau_{\rm s})) & \text{if } \tau \leqslant \tau_{\rm s}, \\ s_{\rm rn} + \tau(1 - s_{\rm rn} - s_{\rm rw}) & \text{if } \tau_{\rm s} \leqslant \tau \leqslant \tau_*, \\ \mathcal{S}_{\rm nw}(\kappa(\tau - \tau_*) + p_*) & \text{if } \tau \geqslant \tau_*, \end{cases} \tag{5.2.6a}$$

$$\mathfrak{p}_{\rm c}(\tau) = \begin{cases} p_{\rm s} + \varsigma(\tau - \tau_{\rm s}) & \text{if } \tau \leqslant \tau_{\rm s}. \\ \mathcal{S}_{\rm nw}^{-1}(s_{\rm rn} + \tau(1 - s_{\rm rn} - s_{\rm rw})) & \text{if } \tau_{\rm s} \leqslant \tau \leqslant \tau_*, \\ \kappa(\tau - \tau_*) + p_* & \text{if } \tau \geqslant \tau_*, \end{cases} \tag{5.2.6b}$$

In the above formulas, $(p_{\rm s}, s_{\rm s}) = (\mathfrak{p}(\tau_{\rm s}), \mathfrak{s}(\tau_{\rm s}))$ is referred later on as the switching point, at which one passes from τ behaving as the saturation to τ behaving as the pressure (recall that Newton's iterations are not sensitive to linear changes of variables). Another switch is incorporated at $(p_*, s_*) = (\mathfrak{p}(\tau_*), \mathfrak{s}(\tau_*))$ to improve Newton's robustness in presence of heterogeneities. The parameter τ_*, such that $s_* = 1 - s_{\rm rw} - \epsilon_* \approx 1 - s_{\rm rw}$, is chosen so that the solution $(p_{{\rm c}K}^n, s_{{\rm nw}K}^n)_{K\in\mathscr{T}}$ to the scheme is always smaller that (p_*, s_*). The parameters κ and ς are chosen so that $\mathfrak{p}$ is C^1, leading to the expressions

$$\kappa = \frac{1 - s_{\rm rn} - s_{\rm rw}}{\mathcal{S}'_{\rm nw}(p_*^+)}, \quad \text{and} \quad \varsigma = \frac{1 - s_{\rm rn} - s_{\rm rw}}{\mathcal{S}'_{\rm nw}(p_{\rm s}^-)}, \tag{5.2.7}$$

where $\mathcal{S}'_{\mathrm{nw}}(p_*^-)$ and $\mathcal{S}'_{\mathrm{nw}}(p_{\mathrm{s}}^+)$ respectively denote the limits of $\mathcal{S}'_{\mathrm{nw}}(p)$ as p tends to p_* and p_{s} from below and above. Then if $\mathcal{S}_{\mathrm{nw}}$ is C^1, so is $\mathfrak{s}_{\mathrm{nw}} = \mathcal{S}_{\mathrm{nw}} \circ \mathfrak{p}_{\mathrm{c}}$. Unfortunately, we lose the concavity of $\mathfrak{p}_{\mathrm{c}}$ because $\mathcal{S}_{\mathrm{nw}}^{-1}$ is convex. An example of parametrized curves $\mathfrak{p}, \mathfrak{s}$ corresponding to Brooks-Corey and van Genuchten-Mualem capillary pressure-(non-wetting) saturation law is shown in Figures 5.3–5.4.

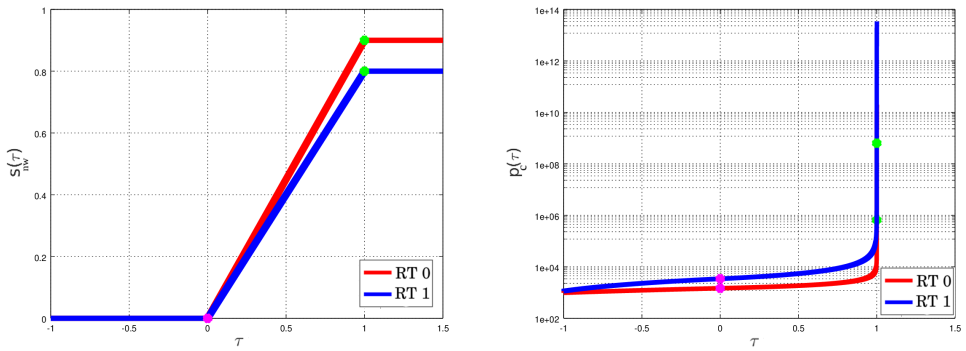


Figure 5.3: Plot of $\mathfrak{p}_{\mathrm{c}}$ and $\mathfrak{s}_{\mathrm{nw}}$ for the Brooks-Corey model, using rock types of Table 5.1.

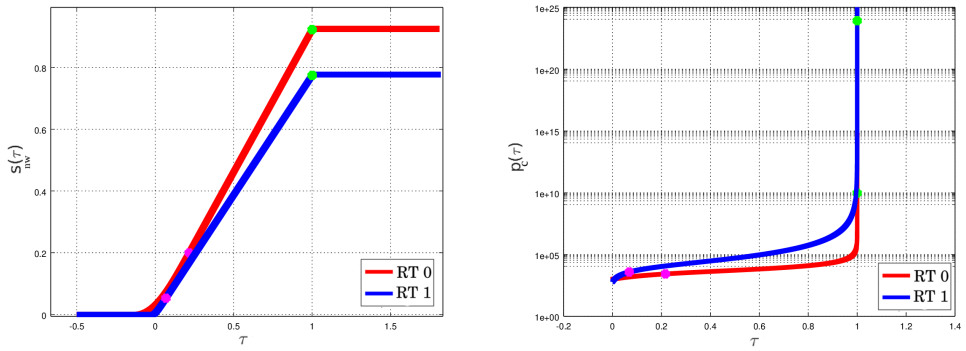


Figure 5.4: Plot of $\mathfrak{p}_{\mathrm{c}}$ and $\mathfrak{s}_{\mathrm{nw}}$ for the van Genuchten-Mualem model, using rock types of Table 5.2.

Applying this parametrization to the previous equations, we obtain the parametrized system

$$m_K \phi_K \frac{\mathfrak{s}_{\mathrm{w}}(\tau_K^n) - \mathfrak{s}_{\mathrm{w}}(\tau_K^{n-1})}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} (F_{\mathrm{w}})_{K\sigma}^n = 0, \tag{5.2.8a}$$

$$m_K \phi_K \frac{\mathfrak{s}_{\mathrm{nw}}(\tau_K^n) - \mathfrak{s}_{\mathrm{nw}}(\tau_K^{n-1})}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} (F_{\mathrm{nw}})_{K\sigma}^n = 0, \tag{5.2.8b}$$

$$\mathfrak{p}_{\mathrm{c}}(\tau_K^n) - [(p_{\mathrm{nw}})_K^n - (p_{\mathrm{w}})_K^n] = 0, \tag{5.2.8c}$$

$$\mathfrak{s}_{\mathrm{nw}}(\tau_K^n) - (\mathcal{S}_{\mathrm{nw}})_K(\mathfrak{p}_{\mathrm{c}}(\tau_K^n)) = 0, \tag{5.2.8d}$$

where the numerical fluxes become

$$(F_{\mathrm{w}})^n_{K\sigma} = \begin{cases} A_\sigma \dfrac{(k_{r,\mathrm{w}})^n_\sigma}{\mu_{\mathrm{w}}} \left[(\vartheta_{\mathrm{w}})^n_K - (\vartheta_{\mathrm{w}})^n_{K\sigma}\right] & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\mathrm{int}} \cup \mathscr{E}^{\mathrm{D}}_{\mathrm{ext}}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} \mathrm{d}t \int_\sigma q^{\mathrm{N}}_{\mathrm{w}} \,\mathrm{d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}^{\mathrm{N}}_{\mathrm{ext}}, \end{cases} \tag{5.2.9a}$$

$$(F_{\mathrm{nw}})^n_{K\sigma} = \begin{cases} A_\sigma \dfrac{(k_{r,\mathrm{nw}})^n_\sigma}{\mu_{\mathrm{nw}}} \left[(\vartheta_{\mathrm{nw}})^n_K - (\vartheta_{\mathrm{nw}})^n_{K\sigma}\right], & \text{if } \sigma \in \mathscr{E}_K \cap (\mathscr{E}_{\mathrm{int}} \cup \mathscr{E}^{\mathrm{D}}_{\mathrm{ext}}), \\ \dfrac{1}{\Delta t^n} \displaystyle\int_{t^{n-1}}^{t^n} \mathrm{d}t \int_\sigma q^{\mathrm{N}}_{\mathrm{nw}} \,\mathrm{d}\gamma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}^{\mathrm{N}}_{\mathrm{ext}}, \end{cases} \tag{5.2.9b}$$

with

$$(\vartheta_{\mathrm{w}})^n_K - (\vartheta_{\mathrm{w}})^n_{K\sigma} = (p_{\mathrm{w}})^n_K - (p_{\mathrm{w}})^n_{K\sigma} + (\psi_{\mathrm{w}})_K - (\psi_{\mathrm{w}})_{K\sigma}, \tag{5.2.10a}$$

$$\begin{aligned}(\vartheta_{\mathrm{nw}})^n_K - (\vartheta_{\mathrm{nw}})^n_{K\sigma} &= (p_{\mathrm{nw}})^n_K - (p_{\mathrm{nw}})^n_{K\sigma} + (\psi_{\mathrm{nw}})_K - (\psi_{\mathrm{nw}})_{K\sigma} \\ &= (p_{\mathrm{w}})^n_K - (p_{\mathrm{w}})^n_{K\sigma} + \mathfrak{p}_{\mathrm{c}}(\tau^n_K) - \mathfrak{p}_{\mathrm{c}}(\tau^n_{K\sigma}) + (\psi_{\mathrm{nw}})_K - (\psi_{\mathrm{nw}})_{K\sigma}\end{aligned} \tag{5.2.10b}$$

and the face mobilities turn into

$$(k_{r,\mathrm{w}})^n_\sigma = \begin{cases} (k_{r,\mathrm{w}})_K(\mathfrak{s}_{\mathrm{nw}}(\tau^n_K)) & \text{if } (\vartheta_{\mathrm{w}})^n_K - (\vartheta_{\mathrm{w}})^n_{K\sigma} \geqslant 0, \\ (k_{r,\mathrm{w}})_{K\sigma}(\mathfrak{s}_{\mathrm{nw}}(\tau^n_{K\sigma})) & \text{if } (\vartheta_{\mathrm{w}})^n_K - (\vartheta_{\mathrm{w}})^n_{K\sigma} < 0. \end{cases} \tag{5.2.11a}$$

$$(k_{r,\mathrm{nw}})^n_\sigma = \begin{cases} (k_{r,\mathrm{nw}})_K(\mathfrak{s}_{\mathrm{nw}}(\tau^n_K)) & \text{if } (\vartheta_{\mathrm{nw}})^n_K - (\vartheta_{\mathrm{nw}})^n_{K\sigma} \geqslant 0, \\ (k_{r,\mathrm{nw}})_{K\sigma}(\mathfrak{s}_{\mathrm{nw}}(\tau^n_{K\sigma})) & \text{if } (\vartheta_{\mathrm{nw}})^n_K - (\vartheta_{\mathrm{nw}})^n_{K\sigma} < 0. \end{cases} \tag{5.2.11b}$$

5.2.2 Treatment of the interface

As in the Richards' equation case, at the interface $\Gamma_{i,j}$, pressures and flux are continuous. More precisely, denote by $p_{\mathrm{w}i}$, $p_{\mathrm{c}i}$ the traces at $(0,T) \times \Gamma_{i,j}$ of the wetting and the capillary pressures $p_{\mathrm{w}|\Omega_i}, p_{\mathrm{c}|\Omega_i}$ in $Q_{i,T}$, and by $v_{\mathrm{w}i}$, $v_{\mathrm{nw}i}$ the traces at $(0,T) \times \Gamma_{i,j}$ of the wetting and non-wetting velocity vectors $v_{\mathrm{w}|\Omega_i}$, $v_{\mathrm{nw}|\Omega_i}$ in $Q_{i,T}$, then the transmission conditions across $\Gamma_{i,j}$ write

$$v_{\mathrm{w}i} \cdot \nu_i + v_{\mathrm{w}j} \cdot \nu_j = 0, \tag{5.2.12a}$$

$$p_{\mathrm{w}i} - p_{\mathrm{w}j} = 0, \tag{5.2.12b}$$

$$v_{\mathrm{nw}i} \cdot \nu_i + v_{\mathrm{nw}j} \cdot \nu_j = 0, \tag{5.2.12c}$$

$$p_{\mathrm{c}i} - p_{\mathrm{c}j} = 0, \tag{5.2.12d}$$

where ν_i (resp. ν_j) denotes the normal to $\Gamma_{i,j}$ outward w.r.t. Ω_i (resp. Ω_j). To approximate these conditions, we can apply the four schemes detailed in §4.3. Concerning the second parametrization introduced in order to enforce the pressure continuity (4.3.12) for Methods C and D, also in this case we need to define monotone functions $\omega_{\sigma,K}, \omega_{\sigma,L}$, with $\omega_{\sigma,K} + \omega_{\sigma,L} > 0$, such that

$$\mathfrak{p}_{\mathrm{c}K}(\omega_{\sigma,K}(\tau)) = \mathfrak{p}_{\mathrm{c}L}(\omega_{\sigma,L}(\tau)), \qquad \forall \sigma \in \Gamma. \tag{5.2.13}$$

Choosing K and L such that $p_{b,K} < p_{b,L}$ for the Brooks-Corey and $\xi_K > \xi_L$ for the van Genuchten-Mualem settings described in (5.2.2)–(5.2.3) respectively, we can choose between the two propositions presented in (4.3.13)–(4.3.14) for functions $\omega_{\sigma,K}, \omega_{\sigma,L}$. The form of the first proposition (4.3.13)

$$\omega^1_{\sigma,K}(\tau) = \tau, \quad \omega^1_{\sigma,L}(\tau) = \mathfrak{p}_L^{-1} \circ \mathfrak{p}_K(\tau),$$

does not change for the two-phase flow case. On the contrary, as we have done for the parametrization, a permutation of the two branches is necessary, i.e.

$$\omega^2_{\sigma,K}(\tau) = \begin{cases} \tau + \beta_K - \beta_L & \text{if } \tau \leqslant \beta_L, \\ \mathfrak{p}_K^{-1} \circ \mathfrak{p}_L(\tau) & \text{if } \tau \geqslant \beta_L. \end{cases} \tag{5.2.14a}$$

$$\omega^2_{\sigma,L}(\tau) = \begin{cases} \mathfrak{p}_L^{-1} \circ \mathfrak{p}_K(\tau + \beta_K - \beta_L) & \text{if } \tau \leqslant \beta_L, \\ \tau & \text{if } \tau \geqslant \beta_L. \end{cases} \tag{5.2.14b}$$

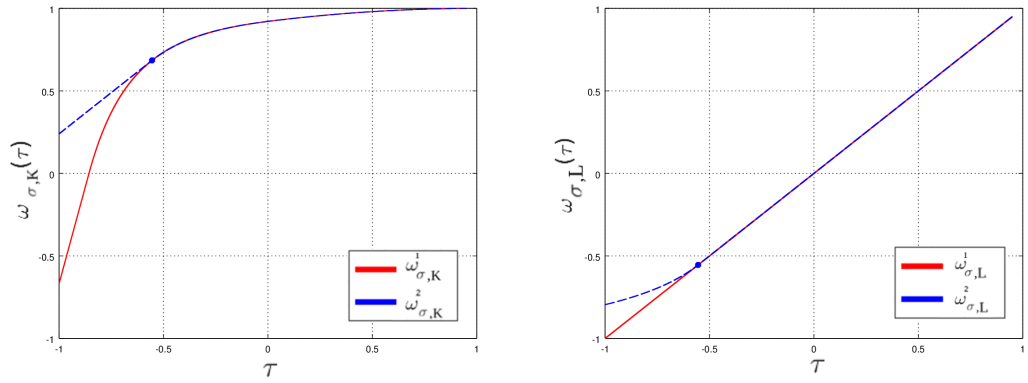


Figure 5.5: Behaviour of $\omega^1_{\sigma,K}(\cdot), \omega^1_{\sigma,L}(\cdot)$ and $\omega^2_{\sigma,K}(\cdot), \omega^2_{\sigma,L}(\cdot)$ functions using the Brooks-Corey model.

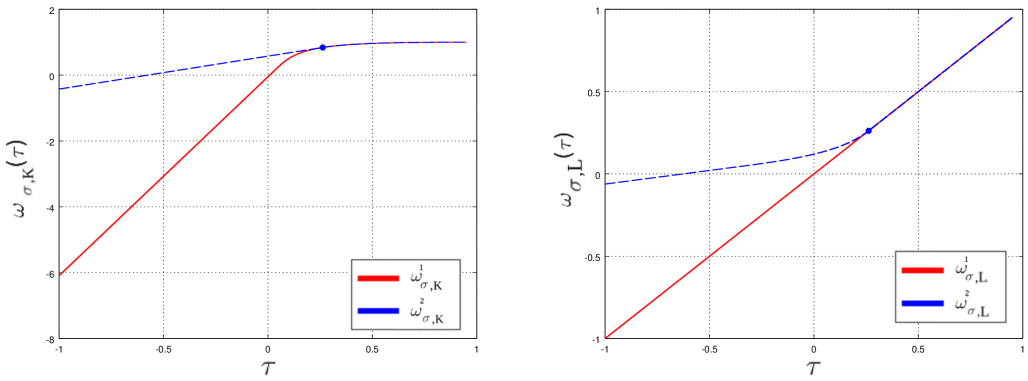


Figure 5.6: Behaviour of $\omega^1_{\sigma,K}(\cdot), \omega^1_{\sigma,L}(\cdot)$ and $\omega^2_{\sigma,K}(\cdot), \omega^2_{\sigma,L}(\cdot)$ functions using the van Genuchten-Mualem model.

The behaviour of $\omega_{\sigma,K}, \omega_{\sigma,L}$ using the Brooks-Corey model and the van Genuchten-Mualem one are reported in Figures 5.5–5.6. As we can notice in the plot, the first and the second propositions overlap for $\tau > \beta_L$ for both models but, for $\tau < \beta_L$, functions behave differently. Considering the Brooks-Corey model, this difference is not relevant because τ remains non-negative all along the simulations. It is no longer the case with the van Genuchten-Mualem model for which the value of β_L is positive. Nevertheless the functions slope for $\tau \in [0, \beta_L]$ remains finite and limited in both cases. Since we are not interested in performing simulations with steep capillary pressure curves,

we postulate that both proposition yield similar numerical behaviors. For sake of simplicity, we will employ just the first proposition, namely $\omega^1_{\sigma,K}$, $\omega^1_{\sigma,L}$.

Remark 5.2.1. *Regarding the resolution of the obtained system considering Method D, in this chapter we only consider the numerical strategy based on the elimination of the face unknowns via the Schur complement (the so-called Method D_1). Consequently, from now on, we remove the index* 1 *and we will just denote it as Method D. The resolution of the system based on the face unknowns elimination thanks to a bisection method is a perspective (see §6.2.2).*

5.3 Numerical validation

We now validate the different approaches we have presented in §5.2.2 to treat the transmission conditions across interfaces when solving the two-phase Darcy flows system in heterogeneous domain via different numerical tests.

5.3.1 CO_2 injection in geological formation

This test is a simplified simulation of the CO_2 physical trapping in a geological formation.

5.3.1.1 Description of the test case

We consider a two-dimensional layered domain $\Omega = [0, 800] \times [-600, 0]$ (in meters) made up of two rock types denoted by RT0 and RT1 respectively, RT1 (which plays the role of cap rock) being less permeable than RT0. Details are reported in Table 5.1 and 5.3. As wetting and non-wetting phases we consider water and CO_2, characterized by the parameters reported in Table 5.4.

	$\lambda[\mathrm{m}^2]$	ϕ
RT0	10^{-11}	0.2
RT1	10^{-13}	0.2

Table 5.3: Parameters of relative permeability and porosity used for RT0 and RT1 with the Brooks-Corey model.

	$\varrho\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$	$\mu\,[\mathrm{Pa}\cdot\mathrm{s}]$
Water	1000	10^{-3}
CO_2	1.795	$1.495\cdot 10^{-3}$

Table 5.4: Parameters used for the wetting (water) and non-wetting phase (CO_2).

The domain Ω is partitioned into three connected subdomains: $\Omega_1 = [0, 800] \times [-200, 0]$, $\Omega_2 = [0, 800] \times [-400, -200]$ and $\Omega_3 = [0, 800] \times [-600, -400]$ (in meters), as depicted in Figure 5.7. The subdomains Ω_1, Ω_3 are characterized by RT0; Ω_2 by RT1.

Starting from an initially water saturated domain, $s^0_{\mathrm{nw}} = 10^{-6}$, CO_2 is leaking into the domain through the bottom boundary $\Gamma^{\mathrm{D}}_1 = \{(x, y) \,|\, y = -600\,\mathrm{m}\}$ on which we impose the boundary conditions

$$(p_{\mathrm{c}})^{\mathrm{D}}_1 = 10^7\,\mathrm{Pa}, \qquad (p_{\mathrm{w}})^{\mathrm{D}}_1 = 0\,\mathrm{Pa}. \tag{5.3.1}$$

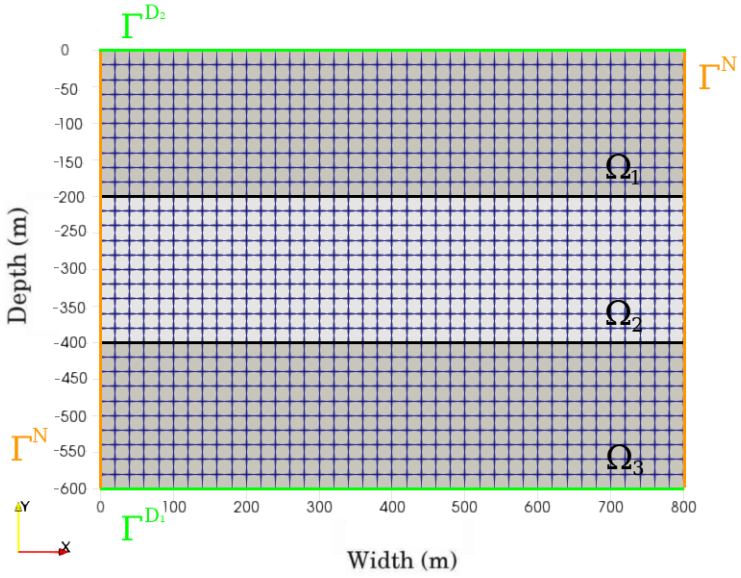


Figure 5.7: Simulation domain $\Omega = [0\text{m}, 800\text{m}] \times [-600\text{m}, 0\text{m}]$.

On the top boundary of the domain $\Gamma_2^{\text{D}} = \{(x, y) \,|\, y = 0\,\text{m}\}$ we set

$$(s_{\text{nw}})_2^{\text{D}} = 0, \qquad (p_{\text{w}})_2^{\text{D}} = 0\,\text{Pa}. \tag{5.3.2}$$

On the lateral boundaries, no-flux boundary conditions are fixed. With these data, CO_2 enters the domain by the bottom and moves upward, whereas water moves first downward then upward. The simulation lasts $T = 116$ days and we adopt an adaptive time-stepping strategy with $\Delta t^0 = 10\,\text{s}$. We take

$$\Delta t^{n+1} = \begin{cases} \min(\Delta t_{\max},\, 1.2\Delta t^n) & \text{for a successful time-step,} \\ \max(\Delta t_{\min},\, 0.5\Delta t^n) & \text{otherwise.} \end{cases}$$

In the latter case, for $\Delta t = \Delta t_{\min}$, the simulation stops. We choose $\Delta t_{\min} = \Delta t^0$ and $\Delta t_{\max} = T \cdot 10^{-2}$. We set $N_{\max} = 20$ as the maximal number of Newton's iterations. The simulation is performed on three structured squared meshes of different resolutions: 20×15, 40×30 and 80×60 cells. It allows us to study the evolution of the solution error measured using the $L^2(\Omega)$-norm of the relative difference at final time between the saturations obtained on a given mesh and the ones computed with Method B and a mesh of resolution 160×120. For this test the classical Brooks-Corey model (§5.2.2) is considered. In Table 5.5 we report the numerical parameters used.

τ_*	ϵ	δ_B	δ_C
10^{-10}	10^{-11}	10^{-1}	$2 \cdot 10^{-1}$

Table 5.5: Numerical parameters used in the examples.

5.3.1.2 Numerical results

We now analyze the results obtained. In Figure 5.8, we report the evolution of the saturation profile at different times obtained on mesh 40×30 using Method B.

Observing Table 5.6 we can remark that Methods A,B,C and D have essentially the same accuracy on this test case, but they present a different behaviour in terms of computational cost. The method that requires the lowest number of Newton iterations is Method A. Method B requires a slightly higher number of iterations but its cost remains comparable to that of Method A. On the other hand the computational cost increases considerably for Methods C and D. Looking at the evolution of the number of iterations for these methods, we notice that it decreases by refining the mesh. For Method C we guess that this behaviour has a connection with the ratio between the size of a cell and the size of the cell introduced on the interface.

Let us better analyze this phenomenon by rerunning the test for Method B and C using different thickness for the interface cell(s). More precisely, using the 40×30 grid, we have considered the thickness values $\delta_B \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ for Method B and $\delta_C \in \{2 \cdot 10^{-1}, 2 \cdot 10^{-2}, 2 \cdot 10^{-3}\}$ for Method C. In Table 5.7, we report data regarding these simulations. As we expected, we can see that the number of iterations and failures increase considerably when δ_B and δ_C decrease. The $L^2(\Omega)$-norm reported in this table is the relative difference at final time between the saturations obtained using the two coarser values for δ_B (δ_C respectively) and the ones computed with the finer one, always employing Method B (Method C respectively) and the mesh of resolution 40×30. We conclude that the thickness chosen for the tests, reported in the Table 5.5, is a good compromise between solution accuracy and computational cost. The key role played by the added mass to the interface unknowns has also an impact on Method D, which is the only one method without mass on the variable interface. Even if mesh refinements yield smaller number of iterations compared to Methods B and C, we notice that it reports a greater number of failures during simulation than the latter two.

Let us try to isolate the behaviour of the different methods analyzing Figure 5.9. This image shows the evolution of the cumulative number of iterations demanded from the Newton method to converge for the simulation carried out on the mesh 40×30. We have already remarked that Methods A and B have a comparable computational cost. Indeed if we look at Figure 5.10, which report the iterations evolution profile for these two methods, we can see that before the CO_2 flux meets the first barrier between Ω_3 and Ω_2 at $t \approx 1.5 \cdot 10^6$, the two curves overlap. Arrived at this point Method B requires lower time steps which implies an increase of the cumulative number of iterations required to converge. Once the non-wetting phase begins to infiltrate Ω_2, the evolution curves for Method A and B follow the same trend until the end of the simulation. The passage of the barrier is also a difficult step for Methods C and D. In Figure 5.9 we see that the number of cumulated iterations increase at $t \approx 1.5 \cdot 10^6$. After the two methods have a different behaviour. Thanks to Figure 5.11 we see that, once the CO_2 flux has crossed the barrier, Method C still encounters some difficulties in converging. Nevertheless it behaves better than Method D. Indeed the latter still encounters important difficulties in converging after the passage of the barrier requiring an average time step around $\Delta t \approx 5000\,\mathrm{s}$, almost one twentieth of the maximum time step. At the end of the simulation the time step increases a little bit but remains lower than the one employed by Method B at the same time.

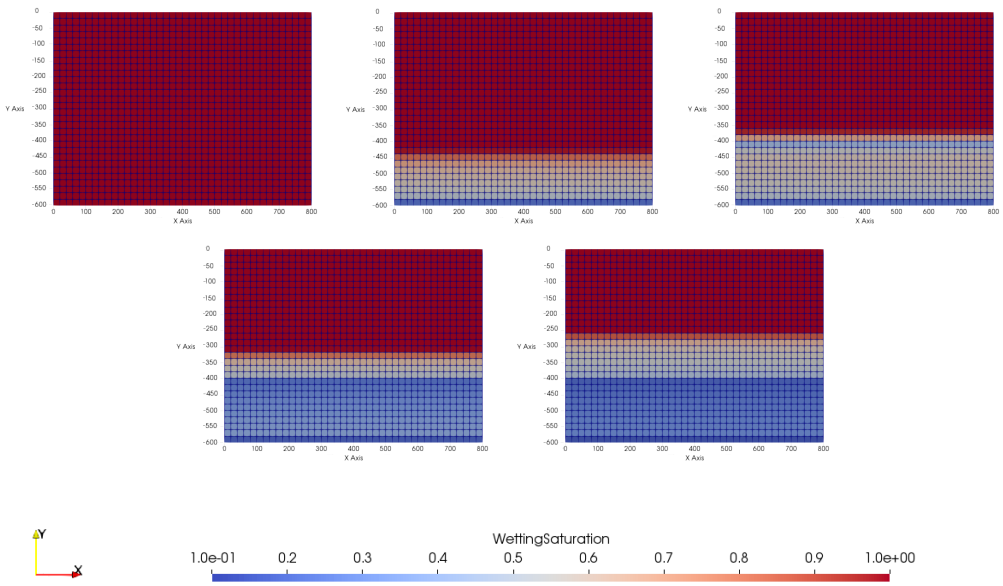


Figure 5.8: Evolution of the saturation profile for $t \in \{0\,\text{s}, 1.14732 \cdot 10^6\,\text{s}, 2.45023 \cdot 10^6\,\text{s}, 5.45695 \cdot 10^6\,\text{s}, 1.00224 \cdot 10^7\,\text{s}\}$ using the Brooks-Corey model, Method B and the 40×30 cells mesh.

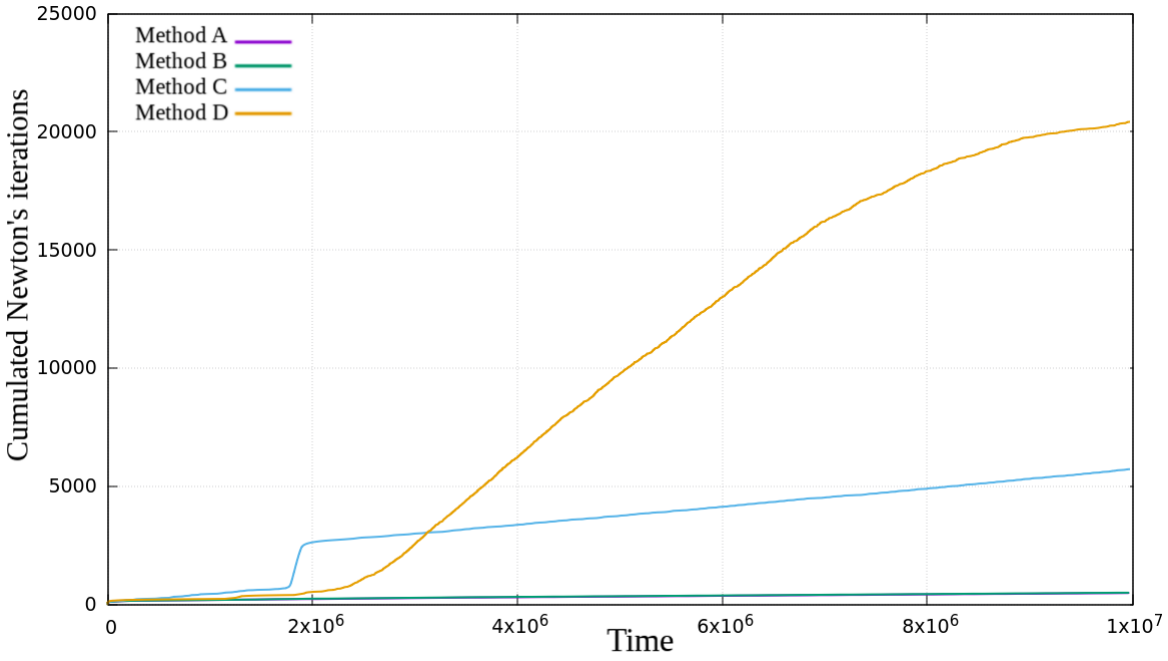


Figure 5.9: Evolution of the average Newton's convergence rate during time iterations for simulation with the Brooks-Corey model on 40×30 cells mesh.

Method A	20×15	40×30	80×60
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	6.24516e-2	3.77878e-2	1.70673e-2
Rate of convergence	–	0.72	1.15
# total iterations	473	479	563
# avg iterations	3	3	3
# max iterations	5	5	5
# failures	0	0	0
Method B	20×15	40×30	80×60
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	6.66594e-2	4.11436e-2	1.80111e-2
Rate of convergence	–	0.72	1.2
# total iterations	502	494	581
# avg iterations	3	3	3
# max iterations	7	8	8
# failures	2	0	1
Method C	20×15	40×30	80×60
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	6.5432e-2	3.93114e-2	1.69376e-2
Rate of convergence	–	0.76	1.21
# total iterations	4005	3507	2973
# avg iterations	8	7	6
# max iterations	20	20	20
# failures	111	111	108
Method D	20×15	40×30	80×60
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	6.85773e-2	4.29154e-2	2.07104e-2
Rate of convergence	–	0.68	1.05
# total iterations	60602	8873	1472
# avg iterations	3	3	4
# max iterations	19	20	20
# failures	4561	578	62

Table 5.6: Results using the Brooks-Corey model.

Method B	$\delta_B = 10^{-1}$	$\delta_B = 10^{-2}$	$\delta_B = 10^{-3}$
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	3.49746e-3	5.83534e-4	–
# total iterations	494	1642	13435
# avg iterations	3	3	2
# max iterations	8	18	20
# failures	0	124	1423
Method C	$\delta_C = 2 \cdot 10^{-1}$	$\delta_C = 2 \cdot 10^{-2}$	$\delta_C = 2 \cdot 10^{-3}$
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	3.9765e-3	1.33095e-3	–
# total iterations	3507	16977	120808
# avg iterations	7	12	13
# max iterations	20	20	20
# failures	111	34	2373

Table 5.7: Evolution of the required Newton iterations w.r.t. the thickness of the interfaces cells for Methods B and C using the Brooks-Corey model with mesh 40×30.

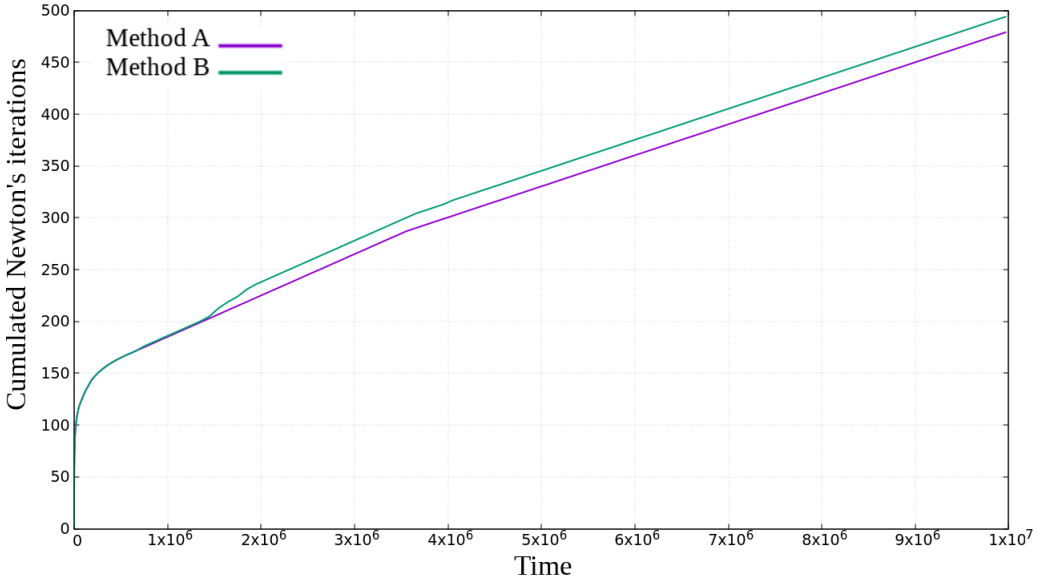


Figure 5.10: Evolution of the average Newton's convergence rate during time iterations for simulation with the Brooks-Corey model on 40×30 cells mesh for Methods A and B.

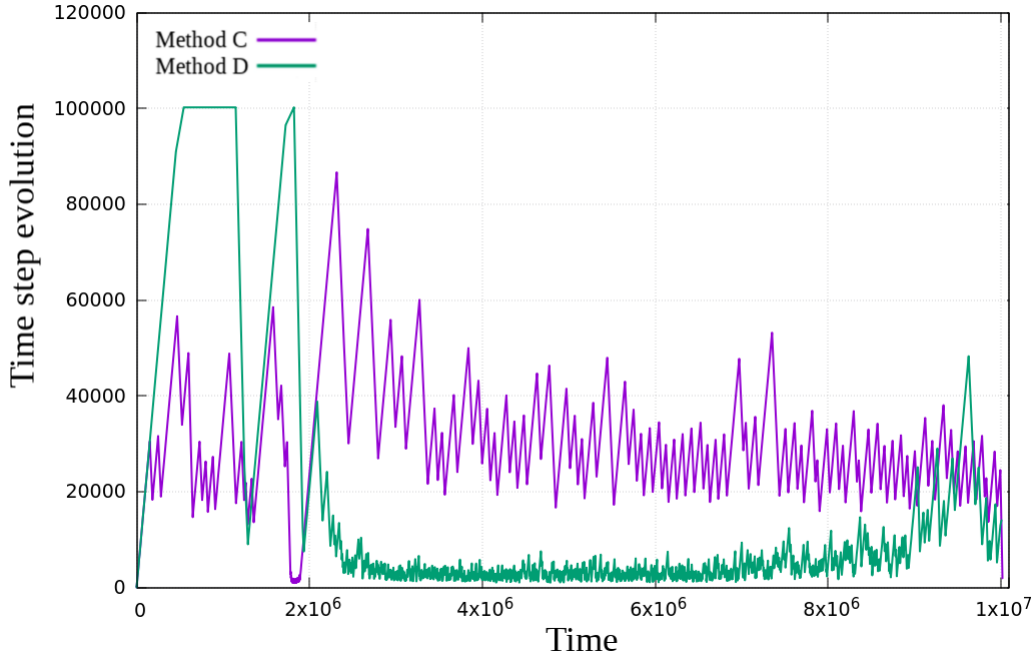


Figure 5.11: Evolution of time step during time iterations for simulation with the Brooks-Corey model on 40×30 cells mesh for Methods C and D.

5.3.2 CO_2 migration towards surface

Thanks to the test case proposed in §5.3.1 we have analyzed the robustness of the proposed strategies to treat heterogeneities in the domain but we have not noticed major improvements in the accuracy of the solution. Thus, to put the different methods to the test and show that, considering the same grid, approaches that introduce an interface-specific treatment provide a more accurate solution than the classical TPFA scheme, we now introduce a second test case. The idea is to consider a case of CO_2 trapping and simulate the migration of CO_2 to the surface through a poorly permeable layer as only an effect of capillarity. Let us detail this test case.

5.3.2.1 Description of the test case

We consider a two-dimensional layered domain $\Omega = [0,1] \times [-0.6, 0]$ (in meters) made up of two rock types denoted by RT0 and RT1 respectively, RT1 (which plays the role of cap rock) being less permeable than RT0. Details are reported in Table 5.1 and 5.3. As wetting and non-wetting phases we consider water and CO_2, characterized by the parameters reported in Table 5.4.

The domain Ω is partitioned into three connected subdomains: $\Omega_1 = [0,1] \times [-0.2, 0]$, $\Omega_2 = [0,1] \times [-0.3, -0.2]$ and $\Omega_3 = [0,1] \times [-0.6, -0.3]$ (in meters), as depicted in Figure 5.12. The subdomains Ω_1, Ω_3 are characterized by RT0; Ω_2 by RT1.

Initially, subdomains Ω_1, Ω_2 are water saturated and Ω_3 is gradually CO_2 saturated. More precise we set the initial capillary pressure profile:

$$p_c^0 = \begin{cases} p_b^{\mathrm{RT0}} & \text{if } -0.2 < z \leqslant 0, \\ p_b^{\mathrm{RT1}} & \text{if } -0.3 < z \leqslant -0.2, \\ p_b^{\mathrm{RT0}} + (\varrho_{\mathrm{w}} - \varrho_{\mathrm{nw}}) g (z + 0.3) & \text{if } -0.6 < z \leqslant -0.3. \end{cases}$$

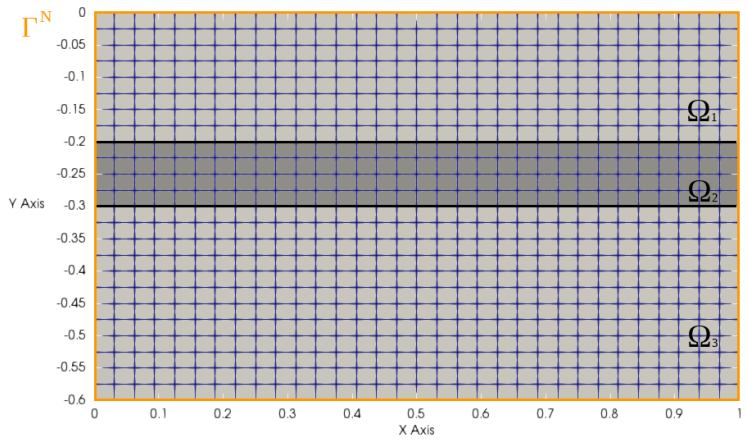


Figure 5.12: Simulation domain $\Omega = [0\text{m}, 1\text{m}] \times [-0.6\text{m}, 0\text{m}]$.

The thickness of Ω_3 has been chosen in order to have, at equilibrium, a certain amount of trapped CO_2, that we denote as h_{acc}, below the cap rock via the relation

$$h_{\text{acc}} = \frac{p_b^{\text{RT1}} - p_b^{\text{RT0}}}{(\varrho_{\text{w}} - \varrho_{\text{nw}})g},$$

where p_b^{RT0}, p_b^{RT1} are the entry pressures of the two lithologies characterizing the domain: rock type 0 and 1 respectively (cf. Table 5.3). Then we take the thickness of Ω_3 slightly greater than the height h_{acc}. On boundaries, no-flux boundary conditions are fixed. The simulation lasts $T = 116$ days and we adopt an adaptive time-stepping strategy with $\Delta t^0 = 100\,\text{s}$. We take

$$\Delta t^{n+1} = \begin{cases} \min(\Delta t_{\max},\, 1.2\Delta t^n) & \text{for a successful time-step,} \\ \max(\Delta t_{\min},\, 0.5\Delta t^n) & \text{otherwise.} \end{cases}$$

In the latter case, for $\Delta t = \Delta t_{\min}$, the simulation stops. We choose $\Delta t_{\min} = \Delta t^0$ and $\Delta t_{\max} = T$. We set $N_{\max} = 20$ as the maximal number of Newton's iterations. The simulation is performed on three structured squared meshes of different resolutions: 8×6, 16×12 and 32×24 cells. A study of the evolution of the solution error measured using the $L^2(\Omega)$-norm of the relative difference at final time between the saturations obtained on a given mesh and the ones computed with Method B and a mesh of resolution 64×48. For this test the classical Brooks-Corey (cf. §5.2.2) model is considered. In Table 5.8 we report the numerical parameters used in the test.

	τ_*	ϵ	δ_B	δ_C
Brooks-Corey	10^{-10}	10^{-12}	10^{-3}	$2 \cdot 10^{-3}$

Table 5.8: Numerical parameters used in the examples.

5.3.2.2 Numerical results

We now analyze the obtained results. In Figure 5.13, we report the evolution of the saturation profile at different times obtained on mesh 40×30 using Method B. Observing Table 5.9 we can immediately remark that the relative L^2 error committed w.r.t. the reference solution using Method

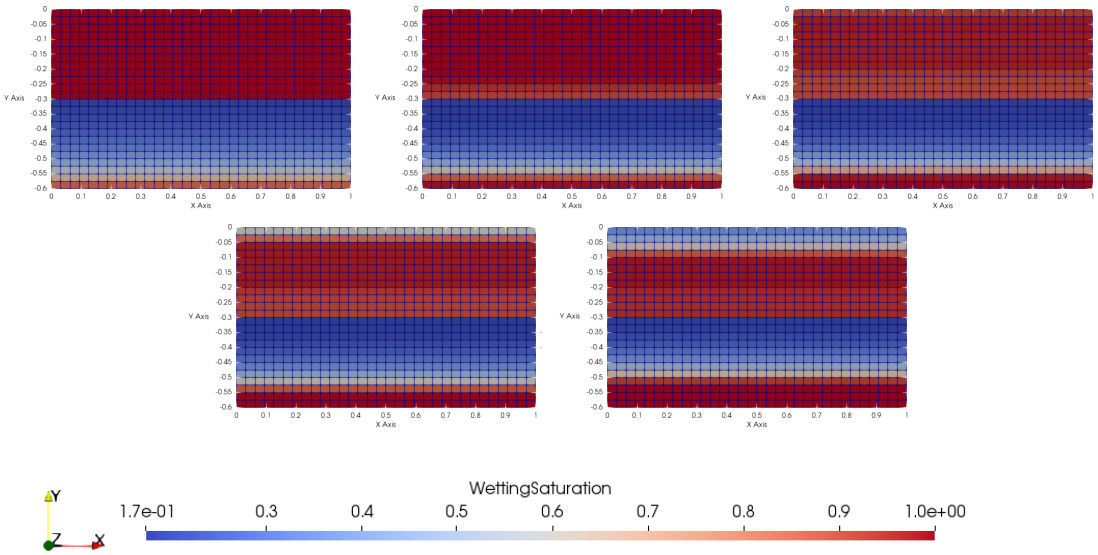


Figure 5.13: Evolution of the saturation profile for $t \in \{0\,\text{s}, 2.04593 \cdot 10^5\,\text{s}, 8.81363 \cdot 10^5\,\text{s}, 1.82813 \cdot 10^6\,\text{s}, 10.022400 \cdot 10^6\,\text{s}\}$ using the Brooks-Corey model, Method B and the 32×24 cells mesh.

A is greater than the one committed by methods which introduce a treatment of the interface. This lack of accuracy is motivated by the fact that the entrapped height h_{acc} for Methods A is calculated from the center of the cells above the interface between Ω_2 and Ω_3, while, for the other methods, it is calculated from the interface. Therefore the amount of trapped CO_2 is underestimated using Method A and a lack of accuracy at the interface level arises. It implies that, if a coarse mesh is considered, Method A is less accurate than the other methods. Looking at Table 5.9 we can observe that method B is the most accurate. Even though method C shows a better accuracy with respect to method A, it is slightly degraded with respect to that of method B.

In the above cited table, data, regarding the simulation using Method D, are not reported because, even if we allow time-steps smaller than the prescribed Δt_{min}, the simulation does not converge. We recall that Method D is the only approach in which an interface variable is added to the system without giving it mass. In the test presented previously (§5.3.1), we had noticed that the addition of mass to the interface variable had an obvious stabilization effect on the simulation (see Table 5.7 and the corresponding analysis performed by refining the thickness of the interface cells). Therefore, the fact that with Method D the simulation does not converge confirms our observation.

Concerning the computational cost, looking at Figure 5.14 we almost observe the same behaviours as in the previous test. More precisely, Methods A and B follow the same trend but Method B requires more iterations at the beginning to converge (passage of the barrier between Ω_2 and Ω_3). It justifies the gap in the cumulated iterations profile between these two methods. Method C is the most expensive one: it suffers from convergence problems during the whole simulation leading to an explosion of the required total number of iterations.

Method A	8×6	16×12	32×24
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	3.40626e-1	1.61846e-1	8.92542e-2
Rate of convergence	$-$	1.07	0.88
# total iterations	251	241	256
# avg iterations	4	4	4
# max iterations	8	7	8
# failures	0	0	0
Method B	8×6	16×12	32×24
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	2.85001e-2	1.64425e-2	7.98564e-3
Rate of convergence	$-$	0.79	1.04
# total iterations	298	334	272
# avg iterations	5	4	4
# max iterations	8	14	15
# failures	0	4	0
Method C	8×6	16×12	32×24
$\frac{\|s^N - s^N_{\text{ref}}\|_{L^2(\Omega)}}{\|s^N_{\text{ref}}\|_{L^2(\Omega)}}$	3.58215e-2	2.50147e-2	1.51162e-2
Rate of convergence	$-$	0.51	0.73
# total iterations	33020	16007	13950
# avg iterations	16	14	12
# max iterations	20	20	20
# failures	517	281	271

Table 5.9: Results using the Brooks-Corey model.

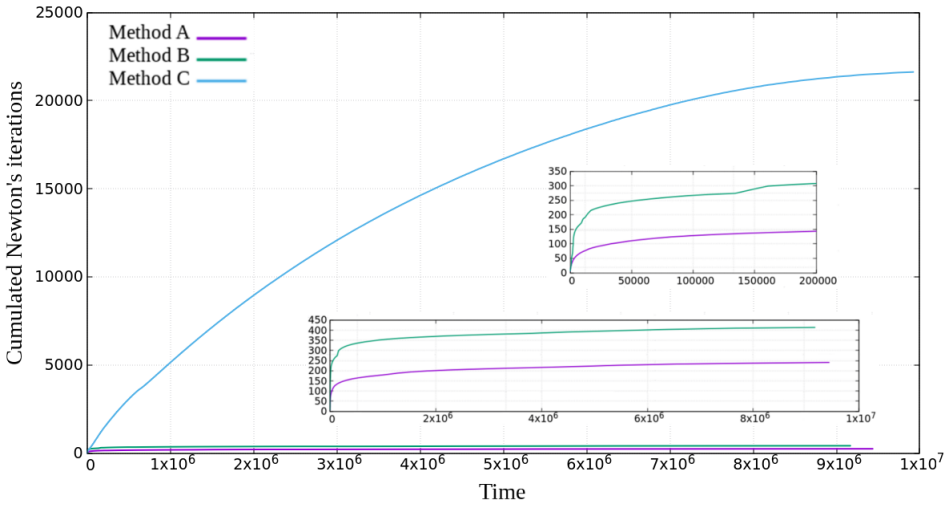


Figure 5.14: Evolution of the average Newton's convergence rate during time iterations for simulation with the Brooks-Corey model on 16×12 cells mesh.

5.3.3 CO_2 migration in layered formation

Let us consider another test case inspired from a test presented in [31]. This test simulates the CO_2 migration in a two-dimensional basin with two barriers.

5.3.3.1 Description of the test case

Let us consider the domain $\Omega = [0, 800] \times [0, 800]$ (in meters) discretized via a uniform mesh of size 40×80. The domain is composed a drain rock sliced by two horizontal barrier layers. More precisely, Ω is partitioned into three connected subdomains: $\Omega_2 = [0, 700] \times [400, 500]$, $\Omega_3 = [100, 800] \times [200, 300]$ and $\Omega_1 = \Omega \setminus (\Omega_2 \cup \Omega_3)$, as depicted in Figure 5.15. The subdomains Ω_2, Ω_3 are characterized by RT1; Ω_1 by RT0. The rock type porosities are $\phi \in \{0.2, 0.2\}$ and their absolute permeabilities are $\lambda \in \{10^{-11}\,\mathrm{m}^2, 10^{-13}\,\mathrm{m}^2\}$ respectively. Details of the constitutive laws' parameters are reported in Table 5.1 and 5.2. The reservoir is initially water saturated and the non-wetting phase (the parameters of wetting and non-wetting phases are reported in Table 5.10) is injected through $\Gamma_1^{\mathrm{D}} = \{(x, y) \,|\, y = 0\,\mathrm{m}\}$ via the Dirichlet conditions

$$(p_{\mathrm{c}})_1^{\mathrm{D}} = 10^7\,\mathrm{Pa}, \qquad (p_{\mathrm{w}})_1^{\mathrm{D}} = 0\,\mathrm{Pa}. \tag{5.3.3}$$

On the top boundary $\Gamma_2^{\mathrm{D}} = \{(x, y) \,|\, y = 800\,\mathrm{m}\}$ the Dirichlet conditions

$$(s_{\mathrm{nw}})_2^{\mathrm{D}} = 0, \qquad (p_{\mathrm{w}})_2^{\mathrm{D}} = 0\,\mathrm{Pa} \tag{5.3.4}$$

are imposed. On the other boundaries, no-flux boundary conditions are fixed.

	$\varrho\,[\mathrm{kg} \cdot \mathrm{m}^{-3}]$	$\mu\,[\mathrm{Pa} \cdot \mathrm{s}]$
wetting phase	1000	10^{-3}
non-wetting phase	1.795	$1.495 \cdot 10^{-3}$

Table 5.10: Parameters used for the wetting and non-wetting phases.

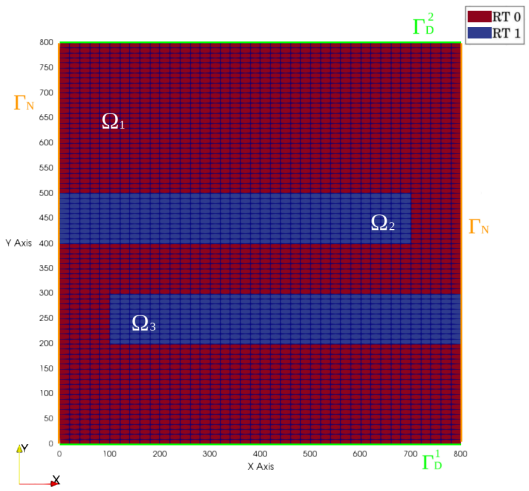


Figure 5.15: Configuration of the domain for the non-wetting injection test case.

The simulation lasts $T = 116$ days and an adaptive time-stepping strategy is adopted using: $\Delta t_{\min} = 1\,\text{s}$, $\Delta t_{\max} = T$. Simulation starts with $\Delta t = \Delta t^0$, then for $n \geqslant 0$,

$$\Delta t^{n+1} = \begin{cases} \min(\Delta t_{\max},\, 1.2\Delta t^n) & \text{for a successful time-step,} \\ \max(\Delta t_{\min},\, 0.5\Delta t^n) & \text{otherwise.} \end{cases}$$

In the latter case, for $\Delta t = \Delta t_{\min}$, the simulation stops. $N_{\max} = 30$ is taken as maximal number of Newton's iterations.

	Δt^0	s_{nw}^0	ϵ_*	ϵ	δ_B	δ_C
Brooks-Corey	10	10^{-6}	10^{-10}	10^{-12}	10^{-1}	$2 \cdot 10^{-1}$
van Genuchten-Mualem	100	$5 \cdot 10^{-5}$	10^{-8}	10^{-12}	10^{-1}	$2 \cdot 10^{-1}$

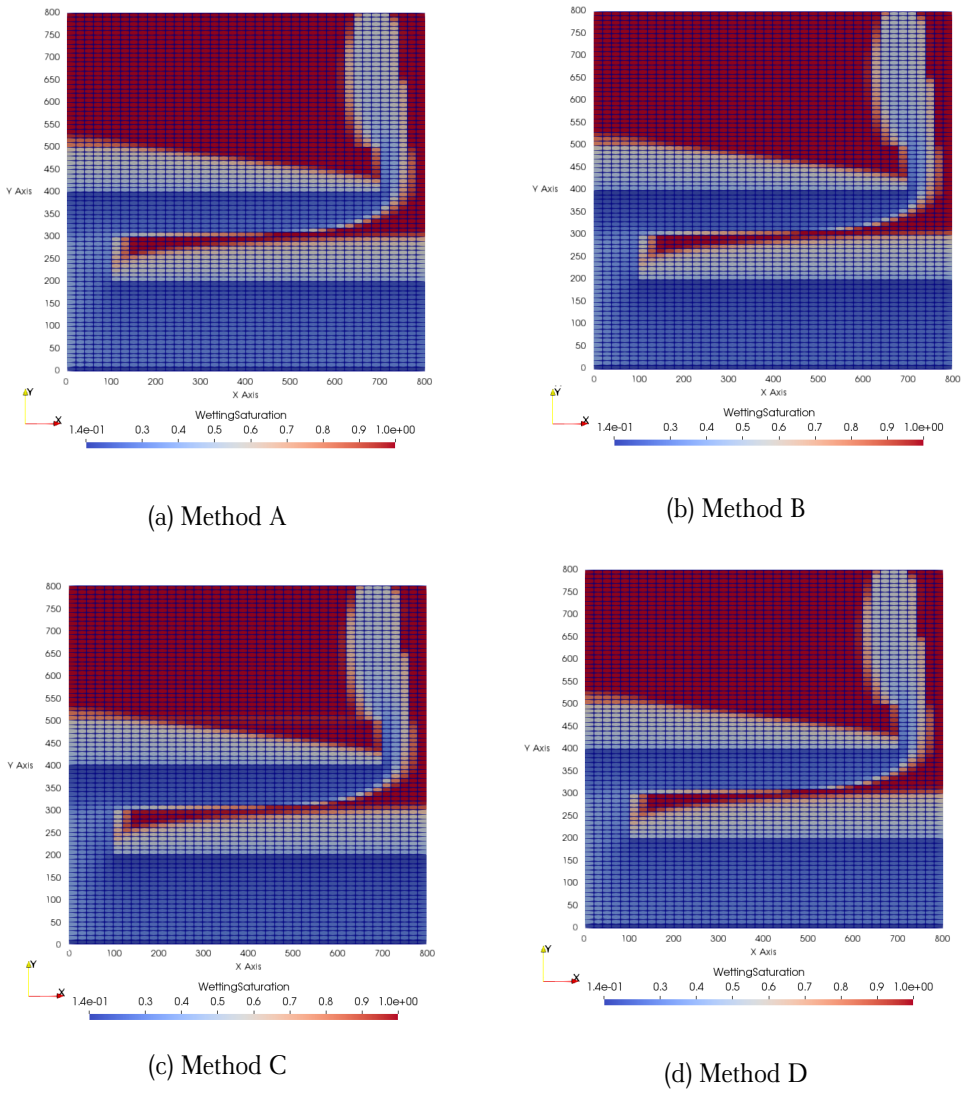
Table 5.11: Numerical parameters used in the examples.

5.3.3.2 Test using the Brooks-Corey model

Let us perform the simulation applying the Brooks-Corey model using, for the lithologies, values in Table 5.1. Figure 5.16 shows the saturation profile at the final time using the four methods. To better visualize the differences between the four profiles, cross sections along the y-axis at the extremities of the two barriers and at the middle of the x-axis are provided in Figure 5.17.

As expected, the various profiles show differences mainly at the barrier level. Leaving aside the areas where the flow can contour the end of a barrier, we can see that in Ω_3 (see Figures 5.17b, 5.17c) the amount of CO_2 that penetrates the barrier is higher with Method A than with the other methods. This is because method A does not have a specific treatment for the barrier: if the mesh is not fine enough, the effect of the capillary barrier is weakened. On the contrary, with the same mesh, the other approaches allow to simulate more accurately the capillary barrier by introducing interface unknowns. Because of the stronger capillary effect with Methods B, C, D, the CO_2 tends to go around the barrier rather than through it. This explains why in Figure 5.17a a higher amount of CO_2 at the extremity of the Ω_3 barrier is obtained with these methods rather with Method A. These comments are also valid for the barrier Ω_2 too.

From the computational-cost point of view, the highlighted behavior of the various methods is consistent with what has been observed in previous tests (cf. Table 5.12). More precisely, Method B requires a slightly larger number of iterations than Method A. But, as shown in Figure 5.18, the evolution of the cumulative number of Newton iterations follows the same trend. Method C encounters more difficulties in converging when crossing the barrier and then stabilizes its number of iterations. In contrast, Method D struggles to converge during the whole simulation.



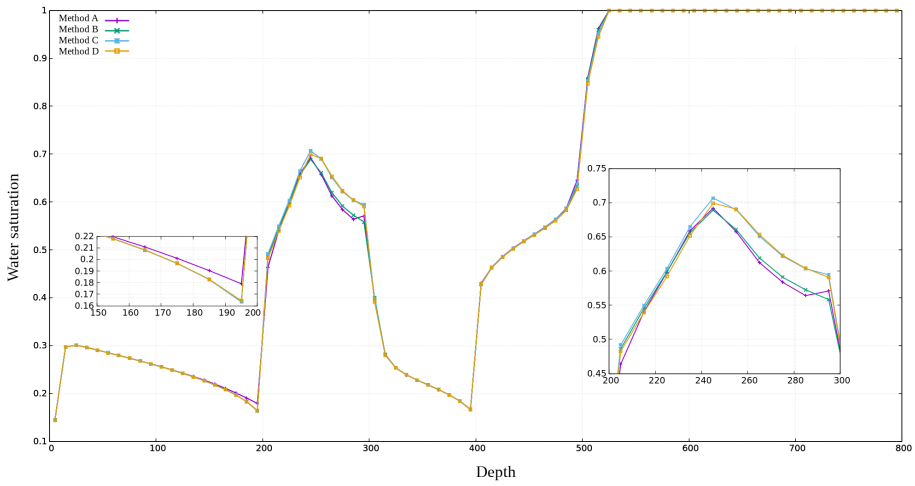
(a) Method A

(b) Method B

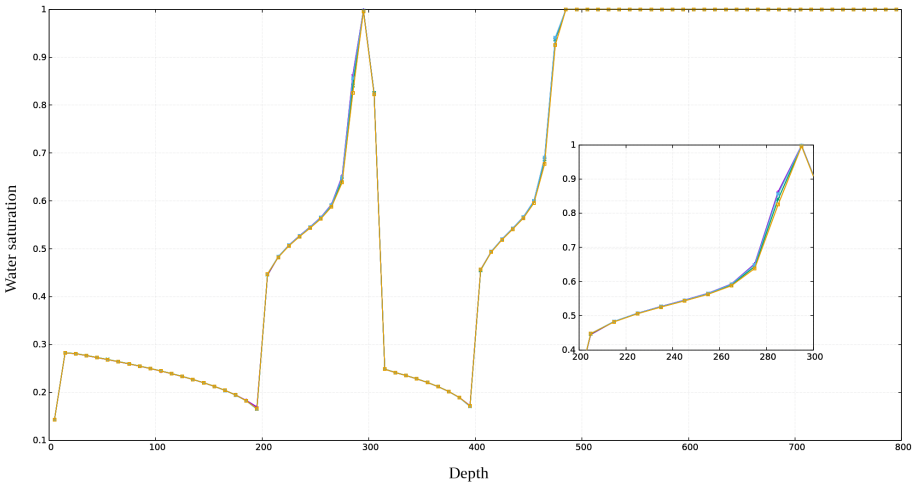
(c) Method C

(d) Method D

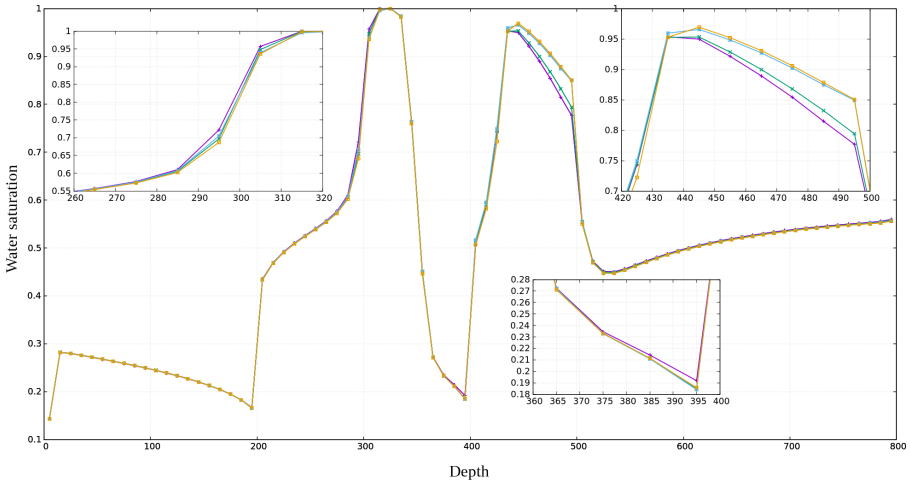
Figure 5.16: Wetting saturation profile at final time of the non-wetting injection simulation with the Brooks-Corey model using Methods A, B, C and D.



(a) Cross-section at $x = 110\,\mathrm{m}$



(b) Cross-section at $x = 390\,\mathrm{m}$



(c) Cross-section at $x = 690\,\mathrm{m}$

Figure 5.17: Vertical cross-sections of the wetting saturation profile at final time of the simulation $t = 116$ days using the Brooks-Corey model.

	# total iterations	# avg iterations	# max iterations	# Newton's failures
Method A	3121	4	28	185
Method B	4354	4	19	262
Method C	20011	16	30	301
Method D	31397	3	29	2398

Table 5.12: Statistics on the required Newton's iterations to converge for test with the Brooks-Corey model.

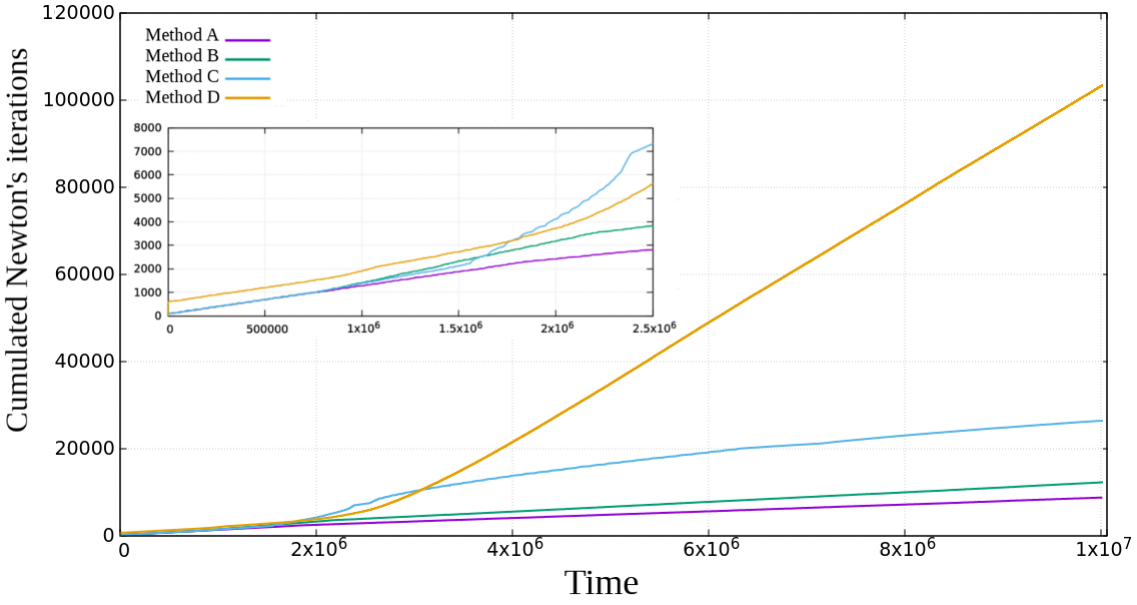
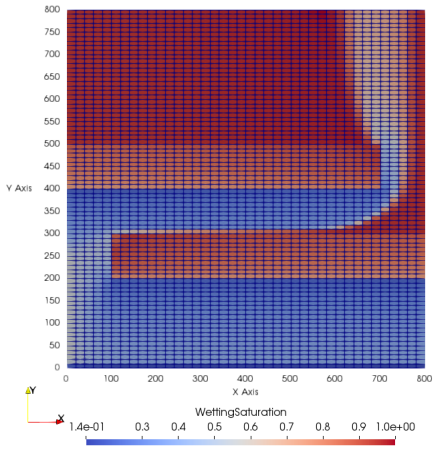


Figure 5.18: Evolution of the average Newton's convergence rate during time iterations for simulation with the Brooks-Corey model.

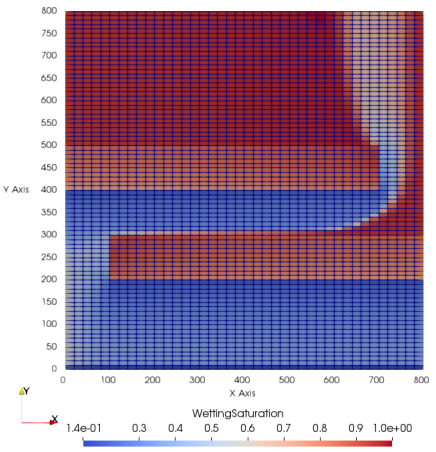
5.3.3.3 Test using the van Genuchten-Mualem model

Let us now perform simulations with the van Genuchten-Mualem model. Lithology data are given in Table 5.2. Figure 5.19 shows the saturation profile at the final time using Methods A,B,C. The lack of results for method D is related to Newton's algorithm which could not converge even by using a time step smaller than $\Delta t_{\min}$. This problem of robustness has already been found out when analyzing the results of the previous tests and the explanations, we then provided, here still apply.

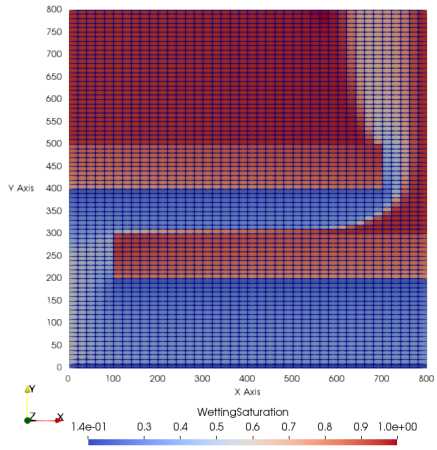
To better visualize the differences between the three profiles, cross sections along the y-axis at the extremities of the two barriers and at the middle of the x-axis are provided in Figure 5.20. We can notice that the saturation profiles obtained with Methods B and C almost completely overlap and correspond to the saturation profile obtained using Method A. Indeed, going back to the results obtained for the filling test case with Richards equation in §4.4.1.1, methods A, B and C almost exhibit the same accuracy. Considering the computational cost and observing Table 5.13 and Figure 5.21, the methods behave in the same way as in the previous tests.



(a) Method A

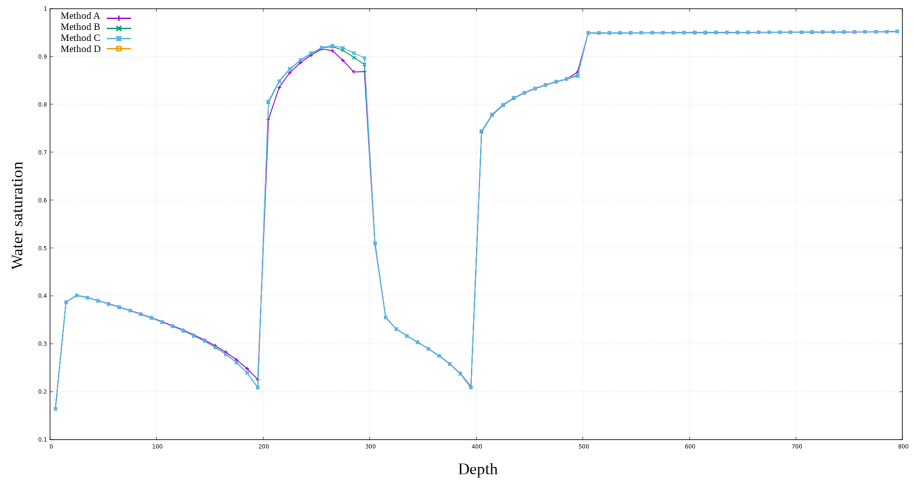


(b) Method B

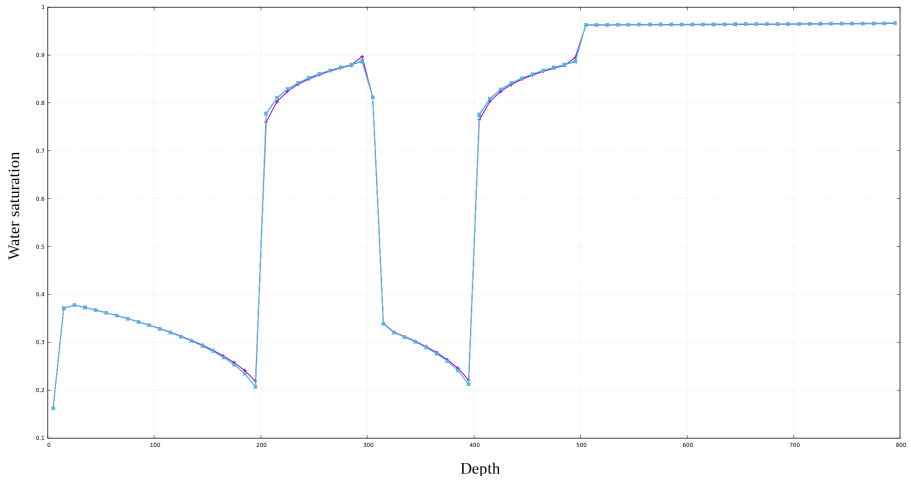


(c) Method C

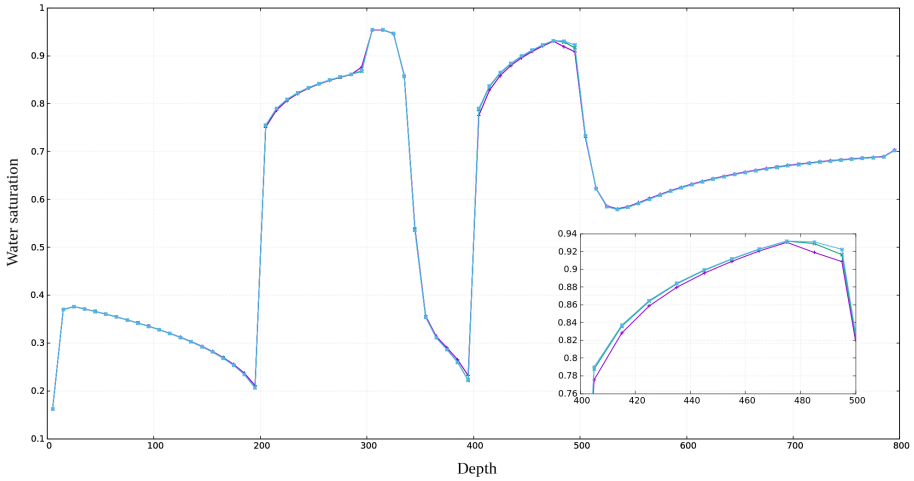
Figure 5.19: Wetting saturation profile at final time of the non-wetting injection simulation with the van Genuchten-Mualem model using Methods A, B, C.



(a) Cross-section at $x = 110\,\mathrm{m}$



(b) Cross-section at $x = 390\,\mathrm{m}$



(c) Cross-section at $x = 690\,\mathrm{m}$

Figure 5.20: Vertical cross-sections of the wetting saturation profile at final time of the simulation $t = 116$ days using the van Genuchten-Mualem model.

	# total iterations	# avg iterations	# max iterations	# Newton's failures
Method A	3531	4	22	195
Method B	4599	4	21	256
Method C	14923	14	30	267

Table 5.13: Statistics on the required Newton's iterations to converge for test with the van Genuchten-Mualem model.

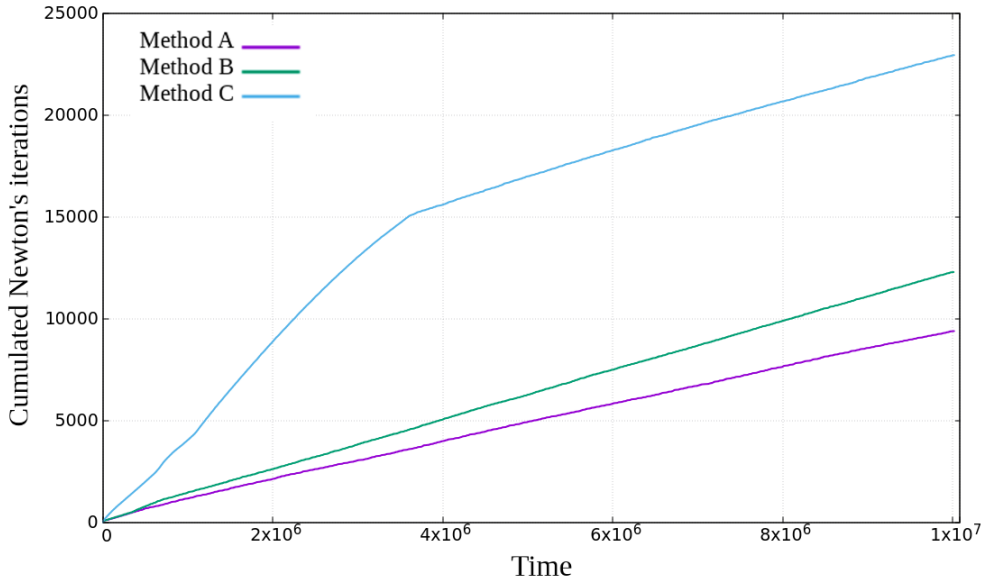


Figure 5.21: Evolution of the average Newton's convergence rate during time iterations for simulation with the van Genuchten-Mualem model.

5.4 Overall method evaluation

We conclude this chapter by drawing conclusions on the behavior of methods B, C and D compared to method A. During the previous tests, it turned out that methods B and C, which add a mass conservation law related to the interface variables are more robust than method D which only uses the flux conservation. For some cases (§5.3.2–§5.3.3.3), Method D was not able to provide the final solution. Moreover, we have observed (Table 5.7) that the amount of added mass influences Newton's convergence: the more the ratio between the mass of the interface cell and the mass of an internal cell is small, the more Newton iterations are required.

Regarding the accuracy of the provided solutions, the test presented in §5.3.1 has shown that a specific treatment of the transmission conditions at the interface allows for a more accurate solution even in case of coarse meshes. Without specific treatment, i.e., using Method A, it is necessary to consider a relative fine mesh to obtain the same order of accuracy. From the standpoint of computational times, methods C and D are quite costly, in particular method D. Method B, on the other hand, exhibits a slight surplus in computational cost compared to Method A but can give more accurate solutions. It therefore seems to be the best option for solving two-phase Darcy flows in heterogeneous domains. Moreover, it entails moderate changes in existing codes in terms of implementation compared to methods C and D.

Chapter 6

Conclusion and perspectives

6.1 Summary of key results

In response to the numerical difficulties stated in §1.3.1, we have conducted works along two lines of research, which have given rise to several presentations at conferences. Besides the FVCA-9 proceeding [18], two scientific publications [19, 20] have been submitted. The first one [20] has been accepted and is about to appear in *ESAIM: Mathematical Modelling and Numerical Analysis*. Below we summarize the most salient results.

6.1.1 Improvement of robustness for Newton's method

As underlined in §1.3.1.1, the numerical resolution of the Richards equation is made challenging not only by the fact that it is a nonlinear, degenerate elliptic-parabolic partial differential equation, but also by a bad choice of the primary unknown for the resolution, which leads to ill-conditioning.

To overcome these difficulties, we have adopted an approach based on a reformulation of the variable switch technique as a parametrization of the graph $\{p, \mathcal{S}(p)\}$ proposed by Brenner and Cancès [29]. Keeping in mind realistic closure laws for which analytical calculation is no longer feasible, we have expanded their idea beyond the Kirchhoff transform-saturation formulation and have primarily focused on the pressure-saturation formulation. Furthermore, to deal with the numerical difficulties introduced by the constitutive laws (blow-up in $k_r'(s)$ for the van Genuchten-Mualem model and kink in $\mathcal{S}(p)$ for the Brooks-Corey model), we have enriched the Newton method specific features to handle such issues.

Numerical tests presented in chapter §2 validated the potentiality of the new parametrization technique. This allows the Richards equation to be solved without caring about the choice of the primary unknown and without any convergence problems due to the selected constitutive laws or the degeneracy of the equation.

6.1.2 Improvement of accuracy for heterogeneous domains

In §1.3.1.2, we highlighted the additional difficulties in the numerical resolution caused by an strong contrast in the constitutive laws' parameters at interface between different lithologies. Using the classical upstream mobility TPFA scheme without any specific treatment for heterogeneities, not only we observe a lack of the accuracy in the predicted results, but also the order of convergence is degraded. The discontinuities in the capillary pressure function between different media are at

the basis of the phenomenon of the capillary barrier, which has a crucial role for flows in porous media. These is why the introduction of specific treatements for interfaces is so important.

In chapter §3, we proved that standard upstream mobility finite volume schemes for variable saturated porous media flows still converge in highly heterogeneous contexts without any specific treatment of the rock type discontinuities. The scheme is indeed shown to satisfy some energy stability which provides enough a priori estimates to carry out its numerical analysis. First, the existence of a unique solution to the nonlinear system stemming from the scheme is established thanks to a topological degree argument and from the monotonicity of the scheme. Besides, a rigorous mathematical convergence proof is conducted, based on compactness arguments. No error estimate can then be deduced from our analysis.

Motivated by the need for a dedicated numerical treatment for the interfaces to enhance the accuracy of the solution and its convergence order, we put forward four methods in chapter §4. The basic the idea is to add unknowns to the system in correspondence/on of the interfaces endowing them with mass or not. The different strategies are compared on filling and drainage test cases with standard nonlinearities of Brooks-Corey and van Genuchten-Mualem type, as well as with challenging steep nonlinearities. The numerical experiments show that the proposed methods allow the lost first-order accuracy to be recovered and the accuracy of the solution to be improved in comparison to that of a naive scheme without any specific treatment for heterogeneities.

Finally, in chapter §5, we successfully extended the parametrization technique and the strategies to treat heterogeneities to the Darcy two-phase flow model. Numerical tests show, as expected, an improved solution accuracy using these strategies compared to that obtained using the scheme without specific treatments for heterogeneities. Nevertheless, contrary to the Richards equation, we have observed that the amount of added mass to interface unknowns plays a very important role in the robustness of the method.

6.2 Recommendations for future research

6.2.1 More advanced models and schemes

The immiscible incompressible two-phase system is barely the most affordable representative of a larger catalog of models, in which more and more physical effects are taken into account. It is natural to move into more advanced models in order to improve the prediction quality of the simulations. The first step would be to incorporate compressibility of both phases. This amounts to considering the interior system

$$\partial_t(\phi\rho_\alpha s_\alpha) + \nabla\cdot(\rho_\alpha v_\alpha) = 0, \tag{6.2.1a}$$

$$v_\alpha + \lambda\frac{k_{r,\alpha}(s_\alpha, x)}{\mu_\alpha}(\nabla p_\alpha - \varrho_\alpha g) = 0, \tag{6.2.1b}$$

$$s_{\mathrm{nw}} - \mathcal{S}_{\mathrm{nw}}(p_{\mathrm{nw}} - p_{\mathrm{w}}, x) = 0, \tag{6.2.1c}$$

$$s_{\mathrm{w}} + s_{\mathrm{nw}} = 1, \tag{6.2.1d}$$

$$\rho_\alpha - \widetilde{\rho}_\alpha(p_\alpha) = 0, \tag{6.2.1e}$$

in which the densities ρ_α are now given functions of the pressures p_α. Next in the complexity scale is the compressible two-phase model with a notion of temperature and an energy equation in order to reflect thermal effects. Last but not least, the most realistic one is the multiphase compositional model with a full thermodynamics involving dynamic appearance and disappearance of phases.

Likewise, the TPFA scheme is the least expensive member of a wider family of finite volume methods. It would be interesting to combine our parametrization technique with more sophisticated

schemes such as MPFA (Multi-Point Flux Approximation) [1,2] for more accuracy on general meshes or nonlinear TPFA [126] for the maximum principle.

6.2.2 Bisection method for the two-phase system

In chapter §5, we did not consider Method D_2 introduced in §4.3.4.2 for the heterogeneous case of the Richards equation. In this method, a face unknown on interfaces is inserted and the system obtained is solved via the face unknowns elimination, thanks to a bisection method. More precisely, the idea is to update the cell unknowns $(\tau_K)_{K\in\mathscr{T}}$ at each Newton iteration ℓ via a linear Schur complement system and then update the interface unknowns $(\tau_\sigma)_{\sigma\in\mathscr{E}_\Gamma}$ by solving exactly the nonlinear wetting flux conservation

$$F^\ell_{K,\sigma}(\vartheta_\sigma) + F^\ell_{L,\sigma}(\vartheta_\sigma) = 0,$$

where

$$\vartheta^\ell_K = \mathfrak{p}_K(\tau^\ell_K) - \varrho g \cdot x_K$$

and

$$F^\ell_{K,\sigma}(\vartheta_\sigma) = \frac{m_\sigma}{d_{K,\sigma}}\lambda_K\eta^\ell_{\sigma,K}(\vartheta_\sigma)[\vartheta^\ell_K - \vartheta_\sigma],$$
$$F^\ell_{L,\sigma}(\vartheta_\sigma) = \frac{m_\sigma}{d_{L,\sigma}}\lambda_L\eta^\ell_{\sigma,L}(\vartheta_\sigma)[\vartheta^\ell_L - \vartheta_\sigma],$$

via a bisection method on each interface. While this was possible for Richards' equation, because the wetting flux conservation is monotone w.r.t. ϑ_σ, things became much more intricate when we turned to the two-phase Darcy flow model. Below we formalize some thoughts in order to circumvent the difficulties encountered.

Let us consider and interface $\sigma = K \mid L$ shared by two different rock types. In cell K —resp. L, the flux is characterized by the wetting pressure $(p_\mathrm{w})_K$ —resp. $(p_\mathrm{w})_L$, the capillary pressure $(p_\mathrm{c})_K$ —resp. $(p_\mathrm{c})_L$— and the non-wetting pressure $(p_\mathrm{nw})_K = (p_\mathrm{w})_K + (p_\mathrm{c})_K$ —resp. $(p_\mathrm{nw})_L = (p_\mathrm{c})_L + (p_\mathrm{w})_L$. We look for the pressure values $(p_\mathrm{w})_\sigma, (p_\mathrm{c})_\sigma$ and $(p_\mathrm{nw})_\sigma = (p_\mathrm{w})_\sigma + (p_\mathrm{c})_\sigma$ on the interface to ensure flux continuity for each phase, i.e.,

$$(F_\mathrm{w})_{K\sigma} + (F_\mathrm{w})_{L\sigma} = 0, \tag{6.2.2a}$$
$$(F_\mathrm{nw})_{K\sigma} + (F_\mathrm{nw})_{L\sigma} = 0. \tag{6.2.2b}$$

The precise expression of the wetting and non-wetting flux $F_\mathrm{w}, F_\mathrm{nw}$ is plugged into equations (5.2.9)–(5.2.10). From these equations, we deduce that if $(p_\mathrm{w})_\sigma$ verifies (6.2.2a), then

$$\min((p_\mathrm{w})_K, (p_\mathrm{w})_L) \leqslant (p_\mathrm{w})_\sigma \leqslant \max((p_\mathrm{w})_K, (p_\mathrm{w})_L). \tag{6.2.3}$$

Analogously, if $(p_\mathrm{nw})_\sigma = (p_\mathrm{w})_\sigma + (p_\mathrm{c})_\sigma$ satisfies (6.2.2b), it verifies

$$\min((p_\mathrm{nw})_K, (p_\mathrm{nw})_L) \leqslant (p_\mathrm{nw})_\sigma \leqslant \max((p_\mathrm{nw})_K, (p_\mathrm{nw})_L).$$

Taking advantage of relation (6.2.3), from (6.2.2a) we can explicit $(p_\mathrm{w})_\sigma$ as

$$(p_\mathrm{w})_\sigma = \alpha_{K\sigma}((p_\mathrm{c})_\sigma)(p_\mathrm{w})_K + \alpha_{L\sigma}((p_\mathrm{c})_\sigma)(p_\mathrm{w})_L,$$

where $\alpha_{K\sigma}$ and $\alpha_{L\sigma}$ are Lipschitz functions taking values in $[0,1]$. Both functions are monotone, respectively increasing and decreasing if the wetting motor $(\vartheta_\mathrm{w})_K - (\vartheta_\mathrm{w})_L$ given by (5.1.5a) is

positive, inversely if negative. The obtained value for $(p_{\mathrm{w}})_\sigma$ can be substituted in the non-wetting flux conservation equation (6.2.2b) and we get that the function

$$\Psi_\sigma : (p_{\mathrm{c}})_\sigma \mapsto (F_{\mathrm{nw}})_{K\sigma}\left((p_{\mathrm{c}})_\sigma\right) + (F_{\mathrm{nw}})_{L\sigma}\left((p_{\mathrm{c}})_\sigma\right)$$

is strictly decreasing and Lipschitz. Finally, depending on the sign of the non-wetting motor $(\vartheta_{\mathrm{nw}})_K - (\vartheta_{\mathrm{nw}})_L$ given by (5.1.5b), we can find the range of values in which search the capillary pressure $(p_{\mathrm{c}})_\sigma$ via the bisection method. Moreover, the sign of the fluxes guarantees the existence and uniqueness of $(p_{\mathrm{c}})_\sigma$ in the previously found interval. It would be interesting to test this method and analyze its behavior to see if it brings advantages over the Method D_1.

Bibliography

[1] I. Aavatsmark, T. Barkve, O. Bœ, and T. Mannseth, *Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media*, J. Comput. Phys., 127 (1996), pp. 2–14, https://doi.org/10.1006/jcph.1996.0154.

[2] L. Agélas, C. Guichard, and R. Masson, *Convergence of finite volume MPFA-O type schemes for heterogeneous anisotropic diffusion problems on general meshes*, Int. J. Finite Vol., 7 (2010), https://hal.archives-ouvertes.fr/hal-00340159.

[3] E. Ahmed, *Splitting-based domain decomposition methods for two-phase flow with different rock types*, Adv. Water Resour., 134 (2019), p. 103431, https://doi.org/10.1016/j.advwatres.2019.103431.

[4] E. Ahmed, C. Japhet, and M. Kern, *Space–time domain decomposition for two-phase flow between different rock types*, Comput. Methods Appl. Mech. Engrg., 371 (2020), p. 113294, https://doi.org/10.1016/j.cma.2020.113294.

[5] A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet, *Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy*, ESAIM: M2AN, 52 (2018), pp. 1533–1567, https://doi.org/10.1051/m2an/2017012.

[6] A. H. Alali, F. P. Hamon, B. T. Mallison, and H. A. Tchelepi, *Finite-volume simulation of capillary-dominated flow in matrix-fracture systems using interface conditions*, Comput. Geosci., 25 (2021), pp. 17–33, https://doi.org/10.1007/s10596-020-09982-1.

[7] H. W. Alt and E. DiBenedetto, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 12 (1985), pp. 335–392, http://www.numdam.org/item?id=ASNSP_1985_4_12_3_335_0.

[8] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Zeit., 183 (1983), pp. 311–341, https://doi.org/10.1007/BF01176474.

[9] H. W. Alt, S. Luckhaus, and A. Visintin, *On nonstationary flow through porous media*, Ann. Mat. Pura Appl., 136 (1984), pp. 303–316, https://doi.org/10.1007/BF01773387.

[10] B. Andreianov, K. Brenner, and C. Cancès, *Approximating the vanishing capillarity limit of two-phase flow in multi-dimensional heterogeneous porous medium*, Z. Angew. Math. Mech., 94 (2014), pp. 651–667, https://doi.org/10.1002/zamm.201200218.

[11] B. Andreianov and C. Cancès, *Vanishing capillarity solutions of Buckley-Leverett equation with gravity in two-rocks' medium*, Comput. Geosci., 17 (2013), pp. 551–572, https://doi.org/10.1007/s10596-012-9329-8.

[12] B. Andreianov, C. Cancès, and A. Moussa, *A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs*, J. Funct. Anal., 273 (2017), pp. 3633–3670, https://doi.org/10.1016/j.jfa.2017.08.010.

[13] B. Andreianov and C. Cancès, *A phase-by-phase upstream scheme that converges to the vanishing capillarity solution for countercurrent two-phase flow in two-rock media*, Comput. Geosci., 18 (2014), pp. 211–226, https://doi.org/10.1007/s10596-014-9403-5.

[14] T. Arbogast, M. Juntunen, J. Pool, and M. F. Wheeler, *A discontinuous Galerkin method for two-phase flow in a porous medium enforcing $H(div)$ velocity and continuous capillary pressure*, Comput. Geosci., 17 (2013), pp. 1055–1078, https://doi.org/10.1007/s10596-013-9374-y.

[15] T. Arbogast, M. Obeyesekere, and M. F. Wheeler, *Numerical methods for the simulation of flow in root-soil systems*, SIAM J. Numer. Anal., 30 (1993), pp. 1677–1702, https://doi.org/10.1137/0730086.

[16] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687, https://doi.org/10.1137/S0036142994266728.

[17] V. Baron, Y. Coudière, and P. Sochala, *Comparison of DDFV and DG methods for flow in anisotropic heterogeneous porous media*, Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles, 69 (2014), pp. 673–686, https://doi.org/10.2516/ogst/2013157.

[18] S. Bassetto, C. Cancès, G. Enchéry, and Q. H. Tran, *Robust Newton solver based on variable switch for a finite volume discretization of Richards equation*, in Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples, R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds., vol. 323 of Springer Proceedings in Mathematics & Statistics, 2020, pp. 385–394, https://doi.org/10.1007/978-3-030-43651-3_35.

[19] S. Bassetto, C. Cancès, G. Enchéry, and Q. H. Tran, *On several numerical strategies to solve Richards' equation in heterogeneous media with Finite Volumes*. working paper or preprint, 2021, https://hal.archives-ouvertes.fr/hal-03259026.

[20] S. Bassetto, C. Cancès, G. Enchéry, and Q. H. Tran, *Upstream mobility Finite Volumes for the Richards equation in heterogenous domains*. to appear in M2AN, 2021, https://hal.archives-ouvertes.fr/hal-03109483.

[21] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander, *A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE*, Computing, 82 (2008), pp. 121–138, https://doi.org/10.1007/s00607-008-0004-9.

[22] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander, *A generic grid interface for parallel and adaptive scientific computing. Part I: Abstract framework*, Computing, 82 (2008), pp. 103–119, https://doi.org/10.1007/s00607-008-0003-x.

[23] J. Bear and Y. Bachmat, *Introduction to modeling of transport phenomena in porous media*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[24] D. B. Bennion and S. Bachu, *The impact of interfacial tension and pore size distribution/capillary pressure character on CO_2 relative permeability at reservoir conditions in CO_2-brine systems*, in SPE/DOE Symposium on Improved Oil Recovery, Tulsa, Oklahoma, USA, 04 2006, https://doi.org/10.2118/99325-MS.

[25] R. R. Berg, *Capillary pressures in stratigraphic traps*, AAPG Bulletin, 59 (1975), pp. 939–956, https://doi.org/10.1306/83D91EF7-16C7-11D7-8645000102C1865D.

[26] L. Bergamaschi and M. Putti, *Mixed finite elements and Newton-type linearizations for the solution of Richards' equation*, Int. J. Numer. Meth. Eng., 45 (1999), pp. 1025–1046, https://doi.org/10.1002/(SICI)1097-0207(19990720)45:8<1025::AID-NME615>3.0.CO;2-G.

[27] M. Bertsch, R. Dal Passo, and C. J. van Duijn, *Analysis of oil trapping in porous media flow*, SIAM J. Math. Anal., 35 (2003), pp. 245–267, https://doi.org/10.1137/S0036141002407375.

[28] K. Brenner, *Acceleration of Newton's method using nonlinear Jacobi preconditioning*, in Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples, R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds., Cham, 2020, Springer International Publishing, pp. 395–403, https://doi.org/10.1007/978-3-030-43651-3_36.

[29] K. Brenner and C. Cancès, *Improving Newton's method performance by parametrization: The case of the Richards equation*, SIAM J. Numer. Anal., 55 (2017), pp. 1760–1785, https://doi.org/10.1137/16M1083414.

[30] K. Brenner, C. Cancès, and D. Hilhorst, *Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure*, Comput. Geosci., 17 (2013), pp. 573–597, https://doi.org/10.1007/s10596-013-9345-3.

[31] K. Brenner, N. Chorfi, and R. Masson, *Sequential implicit Vertex Approximate Gradient discretization of incompressible two-phase Darcy flows with discontinuous capillary pressure.* working paper or preprint, Mar. 2021, https://hal.archives-ouvertes.fr/hal-03176081.

[32] K. Brenner, J. Droniou, R. Masson, and E. H. Quenjel, *Total-velocity-based finite volume discretization of two-phase Darcy flow in highly heterogeneous media with discontinuous capillary pressure*, IMA J. Numer. Anal., drab018 (2021), https://doi.org/10.1093/imanum/drab018.

[33] K. Brenner, M. Groza, C. Guichard, and R. Masson, *Vertex approximate gradient scheme for hybrid dimensional two-phase Darcy flows in fractured porous media*, ESAIM: M2AN, 49 (2015), pp. 303–330, https://doi.org/10.1051/m2an/2014034.

[34] K. Brenner, M. Groza, L. Jeannin, R. Masson, and J. Pellerin, *Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media*, Comput. Geosci., 21 (2017), pp. 1075–1094, https://doi.org/10.1007/s10596-017-9675-7.

[35] K. Brenner, D. Hilhorst, and H.-C. Vu-Do, *A gradient scheme for the discretization of the Richards equation*, in Finite volumes for complex applications. VII. Elliptic, Parabolic and Hyperbolic problems, J. Fuhrmann, M. Ohlberger, and C. Rohde, eds., vol. 78 of Springer Proc. Math. Stat., Springer, Cham, 2014, pp. 537–545, https://doi.org/10.1007/978-3-319-05591-6_53.

[36] K. Brenner, D. Hilhorst, and H.-C. Vu-Do, *The generalized finite volume SUSHI scheme for the discretization of Richards equation*, Vietnam J. Math., 44 (2016), pp. 557—586, https://doi.org/10.1007/s10013-015-0170-y.

[37] K. Brenner, R. Masson, and E. H. Quenjel, *A robust VAG scheme for a two-phase flow problem in heterogeneous porous media*, in Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples, R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds., Cham, 2020, Springer International Publishing, pp. 565–573, https://doi.org/10.1007/978-3-030-43651-3_53.

[38] K. Brenner, R. Masson, and E. H. Quenjel, *Vertex approximate gradient discretization preserving positivity for two-phase Darcy flows in heterogeneous porous media*, J. Comput. Phys., 409 (2020), p. 109357, https://doi.org/10.1016/j.jcp.2020.109357.

[39] R. H. Brooks and A. T. Corey, *Hydraulic properties of porous media*, Hydrology Paper, 7 (1964), pp. 26–28, https://www.wipp.energy.gov/information_repository/cra/CRA-2014/References/Others/Brooks_Corey_1964_Hydraulic_Properties_ERMS241117.pdf.

[40] C. Cancès, *Nonlinear parabolic equations with spatial discontinuities*, Nonlinear Diff. Equ. Appl., 15 (2008), pp. 427–456, https://doi.org/10.1007/s00030-008-6030-7.

[41] C. Cancès, *Finite volume scheme for two-phase flow in heterogeneous porous media involving capillary pressure discontinuities*, ESAIM: M2AN, 43 (2009), pp. 973–1001, https://doi.org/10.1051/m2an/2009032.

[42] C. Cancès, T. Gallouët, and A. Porretta, *Two-phase flows involving capillary barriers in heterogeneous porous media*, Interf. Free Bound., 11 (2009), pp. 239–258, https://doi.org/10.4171/IFB/210.

[43] C. Cancès and C. Guichard, *Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations*, Math. Comp., 85 (2016), pp. 549–580, https://doi.org/10.1090/mcom/2997.

[44] C. Cancès, F. Nabet, and M. Vohralík, *Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations*, Math. Comp., 90 (2021), pp. 517–563, https://doi.org/10.1090/mcom/3577.

[45] C. Cancès and D. Maltese, *A gravity current model with capillary trapping for oil migration in multilayer geological basins*, SIAM J. Appl. Math., 81 (2021), pp. 454–484, https://doi.org/10.1137/19M1284233.

[46] C. Cancès and M. Pierre, *An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field*, SIAM J. Math. Anal., 44 (2012), pp. 966–992, https://doi.org/10.1137/11082943X.

[47] V. Casulli and P. Zanolli, *A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form*, SIAM J. Sci. Comput., 32 (2010), pp. 2255–2273, https://doi.org/10.1137/100786320.

[48] A. J. Cavanagh and R. S. Haszeldine, *The Sleipner storage site: Capillary flow modeling of a layered CO_2 plume requires fractured shale barriers within the Utsira formation*, Int. J. Greenh. Gas Control, 21 (2014), pp. 101–112, https://doi.org/10.1016/j.ijggc.2013.11.017.

[49] M. A. Celia, E. T. Bouloutas, and R. L. Zarba, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res., 26 (1990), pp. 1483–1496, https://doi.org/10.1029/WR026i007p01483.

[50] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng, *Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis*, ESAIM: M2AN, 37 (2003), pp. 319–338, https://doi.org/10.1051/m2an:2003028.

[51] G. Chavent and J. Jaffré, *Mathematical models and finite elements for reservoir simulation: single phase, multiphase and multicomponent flows through porous media*, vol. 17 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1986.

[52] Z. Chen, *Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution*, J. Diff. Eqs, 171 (2001), pp. 203–232, https://doi.org/10.1006/jdeq.2000.3848.

[53] Z. Chen and R. E. Ewing, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology*, SIAM J. Numer. Anal., 34 (1997), pp. 2228–2253, https://doi.org/10.1137/S0036142995290063.

[54] Z. Chen and R. E. Ewing, *Degenerate two-phase incompressible flow. III. Sharp error estimates*, Numer. Math., 90 (2001), pp. 215–240, https://doi.org/10.1007/s002110100291.

[55] H. Darcy, *Les fontaines publiques de la ville de Dijon: Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau*, V. Dalmont, 1856.

[56] K. Deimling, *Nonlinear functional analysis*, Springer-Verlag, Berlin, 1985.

[57] M. Delshad, S. Thomas, and M. Wheeler, *Modeling CO_2 sequestration using a sequentially coupled 'iterative-impec-time-split-thermal' compositional simulator*, ECMOR 2008 - 11th European Conference on the Mathematics of Oil Recovery, (2008), https://doi.org/10.3997/2214-4609.20146390.

[58] H.-J. G. Diersch and P. Perrochet, *On the primary variable switching technique for simulating unsaturated–saturated flows*, Adv. Water Resour., 23 (1999), pp. 271–301, https://doi.org/10.1016/S0309-1708(98)00057-8.

[59] B. Doyle, B. Rivière, and M. Sekachev, *A multinumerics scheme for incompressible two-phase flow*, Computer Methods in Applied Mechanics and Engineering, 370 (2020), p. 113213, https://doi.org/10.1016/j.cma.2020.113213.

[60] J. Droniou, *Finite volume schemes for diffusion equations: introduction to and review of modern methods*, Math. Models Meth. Appl. Sci., 24 (2014), pp. 1575–1619, https://doi.org/10.1142/S0218202514400041.

[61] J. Droniou and R. Eymard, *The asymmetric gradient discretisation method*, in Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects, C. Cancès and P. Omnes, eds., vol. 199 of Springer Proc. Math. Stat., Cham, 2017, Springer, pp. 311–319, https://doi.org/10.1007/978-3-319-57397-7_24.

[62] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin, *The Gradient Discretisation Method*, vol. 42 of Mathématiques et Applications, Springer International Publishing, 2018, https://doi.org/10.1007/978-3-319-79042-8.

[63] J. Droniou, R. Eymard, T. Gallouet, and R. Herbin, *A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods*, Math. Models Meth. Appl. Sci., 20 (2010), pp. 265–295, https://doi.org/10.1142/S0218202510004222.

[64] J. Droniou, J. Hennicker, and R. Masson, *Numerical analysis of a two-phase flow discrete fracture matrix model*, Numer. Math., 141 (2019), pp. 21–62, https://doi.org/10.1007/s00211-018-0994-y.

[65] G. Enchéry, R. Eymard, and A. Michel, *Numerical approximation of a two-phase flow problem in a porous medium with discontinuous capillary forces*, SIAM J. Numer. Anal., 43 (2006), pp. 2402–2422, https://doi.org/10.1137/040602936.

[66] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, vol. 159, Springer Science & Business Media, 2013, https://doi.org/10.1007/978-1-4757-4355-5.

[67] A. Ern, I. Mozolevski, and L. Schuh, *Discontinuous Galerkin approximation of two-phase flows in heterogeneous porous media with discontinuous capillary pressures*, Comput. Meth. Appl. Mech. Engrg, 199 (2010), pp. 1491–1501, https://doi.org/10.1016/j.cma.2009.12.014.

[68] A. Ern, I. Mozolevski, and L. Schuh, *Discontinuous Galerkin approximation of two-phase flows in heterogeneous porous media with discontinuous capillary pressures*, Comput. Meth. Appl. Mech. Engrg, 199 (2010), pp. 1491–1501, https://doi.org/10.1016/j.cma.2009.12.014.

[69] B. G. Ersland, M. S. Espedal, and R. Nybo, *Numerical methods for flows in a porous medium with internal boundary*, Comput. Geosci., 2 (1998), pp. 217–240, https://doi.org/10.1023/A:1011554320427.

[70] R. Eymard and T. Gallouët, *H-convergence and numerical schemes for elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 539–562, https://doi.org/10.1137/S0036142901397083.

[71] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson, *TP or not TP, that is the question*, Comput. Geosci., 18 (2014), pp. 285–296, https://doi.org/10.1007/s10596-013-9392-9.

[72] R. Eymard, T. Gallouët, R. Herbin, M. Gutnic, and D. Hilhorst, *Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies*, Math. Models Meth. Appl. Sci., 11 (2001), pp. 1505–1528, https://doi.org/10.1142/S0218202501001446.

[73] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3), P. G. Ciarlet and J. L. Lions, eds., vol. 7 of Handbook of Numerical Analysis, Elsevier, 2000, pp. 713–1018, https://doi.org/10.1016/S1570-8659(00)07005-8.

[74] R. Eymard, C. Guichard, R. Herbin, and R. Masson, *Gradient schemes for two-phase flow in heterogeneous porous media and richards equation*, Z. Angew. Math. Mech., 94 (2014), pp. 560–585, https://doi.org/10.1002/zamm.201200206.

[75] R. Eymard, M. Gutnic, and D. Hilhorst, *The finite volume method for Richards equation*, Comput. Geosci., 3 (1999), pp. 259–294, https://doi.org/10.1023/A:1011547513583.

[76] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum-engineering scheme*, ESAIM: M2AN, 37 (2003), pp. 937–972, https://doi.org/10.1051/m2an:2003062.

[77] M. S. Fabien, M. G. Knepley, and B. M. Rivière, *A hybridizable discontinuous Galerkin method for two-phase flow in heterogeneous porous media*, Int. J. Numer. Meth. Engrg, 116 (2018), pp. 161–177, https://doi.org/10.1002/nme.5919.

[78] M. W. Farthing and F. L. Ogden, *Numerical solution of Richards' equation: A review of advances and challenges*, Soil Sci. Soc. Amer. J., 81 (2017), pp. 1257–1269, https://doi.org/10.2136/sssaj2017.02.0058.

[79] B. Flemisch, J. Fritz, R. Helmig, and J. Niessner, *DUMUX: a multi-scale multi-physics toolbox for flow and transport processes in porous media*, in ECCOMAS Thematic Conference on Multi-scale Computational Methods for Solids and Fluids, A. Ibrahimbegovic, F. Dias, H. Matthies, and P. Wriggers, eds., 2007, https://www.iws.uni-stuttgart.de/publikationen/hydrosys/paper/cachan_flemisch.pdf.

[80] P. A. Forsyth, *A control volume finite element approach to NAPL groundwater contamination.*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1029–1057, https://doi.org/10.1137/0912055.

[81] P. A. Forsyth, Y. S. Wu, and K. Pruess, *Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media*, Adv. Water Resour., 18 (1995), pp. 25–38, https://doi.org/10.1016/0309-1708(95)00020-J.

[82] K. Gärtner and L. Kamenski, *Why do we need Voronoi cells and Delaunay meshes?*, in Numerical Geometry, Grid Generation and Scientific Computing, V. A. Garanzha, L. Kamenski, and H. Si, eds., Lecture Notes in Computational Science and Engineering, Cham, 2019, Springer International Publishing, pp. 45–60, https://doi.org/10.1007/978-3-030-23436-2_3.

[83] S. E. Gasda, *Numerical models for evaluating CO_2 storage in deep, saline aquifers: Leaky wells and large-scale geological features*, PhD thesis, Princeton University, 2008, http://arks.princeton.edu/ark:/88435/dsp01j098zb09n.

[84] Y. Ghomian, G. A. Pope, and K. Sepehrnoori, *Hysteresis and field-scale optimization of WAG injection for coupled CO_2-EOR and sequestration*, in SPE Symposium on Improved Oil Recovery, Tulsa, Oklahoma, USA, 04 2008, https://doi.org/10.2118/110639-MS.

[85] V. Girault, B. Rivière, and L. Cappanera, *Convergence of a finite element method for degenerate two-phase flow in porous media.* hal-02453608, 2020, https://hal.archives-ouvertes.fr/hal-02453608.

[86] J. Gomes and A. Bera, *A review of CO_2 storage in geological formations emphasizing modeling, monitoring and capacity estimation approaches*, Petrol. Sci., 16 (2019), pp. 1–36, https://doi.org/10.1007/s12182-019-0340-8.

[87] M. Grobe, J. Pashin, and R. Dodge, *Carbon dioxide sequestration in geological media—state of the science*, in Carbon dioxide sequestration in geological media—State of the science, M. Grobe, J. Pashin, and R. Dodge, eds., vol. 59 of AAPG Studies in Geology, American Association of Petroleum Geologists, 2009, pp. 1–2, https://doi.org/10.1306/1371229St591675.

[88] F. P. Hamon, B. T. Mallison, and H. A. Tchelepi, *Implicit hybrid upwinding for two-phase flow in heterogeneous porous media with buoyancy and capillarity*, Comput. Methods Appl. Mech. Engrg., 331 (2018), pp. 701–727, https://doi.org/10.1016/j.cma.2017.10.008.

[89] W. S. Han, B. J. McPherson, P. C. Lichtner, and F. P. Wang, *Evaluation of trapping mechanisms in geologic CO_2 sequestration: Case study of SACROC northern platform, a 35-year CO_2 injection site*, Amer. J. Sci., 310 (2010), pp. 282–324, https://doi.org/10.2475/04.2010.03.

[90] H. Hoteit and A. Firoozabadi, *Numerical modeling of two-phase flow in heterogeneous permeable media with different capillarity pressures*, Adv. Water Resour., 31 (2008), pp. 56–73, https://doi.org/10.1016/j.advwatres.2007.06.006.

[91] J. Hou, J. Chen, S. Sun, and Z. Chen, *Adaptive mixed-hybrid and penalty discontinuous Galerkin method for two-phase flow in heterogeneous media*, J. Comput. Appl. Math., 307 (2016), pp. 262–283, https://doi.org/10.1016/j.cam.2016.01.050.

[92] S. D. Hovorka, S. M. Benson, C. Doughty, B. M. Freifeld, S. Sakurai, T. M. Daley, Y. K. Kharaka, M. H. Holtz, R. C. Trautz, H. S. Nance, L. R. Myer, and K. G. Knauss, *Measuring permanence of CO_2 storage in saline formations: the Frio experiment*, Environmental Geosciences, 13 (2006), pp. 105–121, https://doi.org/10.1306/eg.11210505011.

[93] P. Jenny, H. A. Tchelepi, and S. H. Lee, *Unconditionally convergent nonlinear solver for hyperbolic conservation laws with S-shaped flux functions*, J. Comput. Phys., 228 (2009), pp. 7497–7512, https://doi.org/10.1016/j.jcp.2009.06.032.

[94] J. Jiang and H. A. Tchelepi, *Dissipation-based continuation method for multiphase flow in heterogeneous porous media*, J. Comput. Phys., 375 (2018), pp. 307–336, https://doi.org/10.1016/j.jcp.2018.08.044.

[95] X. Jiang, *A review of physical modelling and numerical simulation of long-term geological storage of CO_2*, Applied Energy, 88 (2011), pp. 3557–3566, https://doi.org/10.1016/j.apenergy.2011.05.004.

[96] M. S. Joshaghani, V. Girault, and B. Riviere, *A vertex scheme for two-phase flow in heterogeneous media*, 2021, https://arxiv.org/abs/2103.03285.

[97] M. S. Joshaghani, B. Rivière, and M. Sekachev, *Maximum-principle-satisfying discontinuous Galerkin methods for incompressible two-phase immiscible flow*, 2021, https://arxiv.org/abs/2106.11807.

[98] R. Juanes, E. J. Spiteri, F. M. Orr Jr., and M. J. Blunt, *Impact of relative permeability hysteresis on geological CO_2 storage*, Water Resour. Res., 42 (2006), p. 12418, https://doi.org/10.1029/2005WR004806.

[99] M. R. Kirkland, R. G. Hills, and P. J. Wierenga, *Algorithms for solving Richards equation for variably saturated soils*, Water Resour. Res., 28 (1992), pp. 2049–2058, https://doi.org/10.1029/92WR00802.

[100] R. A. Klausen, F. A. Radu, and G. T. Eigestad, *Convergence of MPFA on triangulations and for Richards' equation*, Int. J. Numer. Meth. Fluids, 58 (2008), pp. 1327–1351, https://doi.org/10.1002/fld.1787.

[101] D. Kröner and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[102] T. Laurent, A. Michel, E. Tillier, and Y. Le Gallo, *A sequential splitting strategy for CO_2 storage modelling*, in Proceedings of the 10th European Conference on the Mathematics of Oil Recovery, 09 2006, pp. 23–00041, https://doi.org/10.3997/2214-4609.201402512.

[103] Y. Le Gallo, T. Laurent, A. Michel, S. Vidal-Gilbert, and T. Parra, *Long-term flow simulations of CO_2 storage in saline aquifer*, in Proceedings of the GHGT8 conference, Trondheim, Norway, 06 2006, pp. 18–22.

[104] F. Lehmann and P. Ackerer, *Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media*, Transp. Porous Media, 31 (1998), pp. 275–292, https://doi.org/10.1023/A:1006555107450.

[105] J. Leray and J. Schauder, *Topologie et équations fonctionnelles*, Ann. Sci. École Norm. Sup., 51 (1934), pp. 45–78, http://www.numdam.org/article/ASENS_1934_3_51__45_0.pdf.

[106] H. Li, M. W. Farthing, C. N. Dawson, and M. C. T., *Local discontinuous Galerkin approximations to Richards' equation*, Adv. Water. Resour., 30 (2007), pp. 555–575, https://doi.org/10.1016/j.advwatres.2006.04.011.

[107] F. List and F. A. Radu, *A study on iterative methods for solving Richards' equation*, Comput. Geosci., 20 (2016), pp. 341–353, https://doi.org/10.1007/s10596-016-9566-3.

[108] F. Marinelli and D. S. Dunford, *Semianalytical solution to Richards equation for layered porous media*, J. Irrig. Drain. Eng., 124 (1998), pp. 290–299, https://doi.org/10.1061/(ASCE)0733-9437(1998)124:6(290).

[109] D. McBride, M. Cross, N. Croft, C. Bennett, and J. Gebhardt, *Computational modelling of variably saturated flow in porous media with complex three-dimensional geometries*, Int. J. Numer. Meth. Fluids, 50 (2006), pp. 1085–1117, https://doi.org/10.1002/fld.1087.

[110] K. Mitra, T. Köppl, C. Duijn, I. Pop, and R. Helmig, *Fronts in two-phase porous flow problems: effects of hysteresis and dynamic capillarity*, Stud. Appl. Math., 144 (2020), pp. 449–492, https://doi.org/10.1111/sapm.12304.

[111] I. Mozolevski and L. Schuh, *Numerical simulation of two-phase immiscible incompressible flows in heterogeneous porous media with capillary barriers*, J. Comput. Appl. Math., 242 (2013), pp. 12–27, https://doi.org/10.1016/j.cam.2012.09.045.

[112] M. Muskat, R. Wyckoff, and R. Wyckoff, *The Flow of Homogeneous Fluids Through Porous Media*, International Series in Physics, McGraw-Hill Book Company, Incorporated, 1937.

[113] T. D. Ngo, E. Mouche, and P. Audigane, *Buoyant flow of CO_2 through and around a semi-permeable layer of finite extent*, J. Fluid Mech., 809 (2016), pp. 553–584, https://doi.org/10.1017/jfm.2016.684.

[114] A. Nicol, R. Carne, M. Gerstenberger, and A. Christophersen, *Induced seismicity and its implications for CO_2 storage risk*, Energy Procedia, 4 (2011), pp. 3699–3706, https://doi.org/10.1016/j.egypro.2011.02.302.

[115] J. Niessner, R. Helmig, H. Jakobs, and J. E. Roberts, *Interface condition and linearization schemes in the Newton iterations for two-phase flow in heterogeneous porous media*, Adv. Water Resour., 28 (2005), pp. 671–687, https://doi.org/10.1016/j.advwatres.2005.01.006.

[116] J. Nordbotten, M. Celia, and S. Bachu, *Injection and storage of CO_2 in deep saline aquifers: Analytical solution for CO_2 plume evolution during injection*, Transp. Porous Media, 58 (2005), pp. 339–360, https://doi.org/10.1007/s11242-004-0670-9.

[117] F. Otto, *L^1-contraction and uniqueness for unstationary saturated-unsaturated porous media flow*, Adv. Math. Sci. Appl., 7 (1997), pp. 537–553.

[118] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373, https://doi.org/10.1016/j.cam.2003.04.008.

[119] F. A. Radu, K. Kumar, J. M. Nordbotten, and I. S. Pop, *A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities*, IMA J. Numer. Anal., 38 (2017), pp. 884–920, https://doi.org/10.1093/imanum/drx032.

[120] F. A. Radu, I. S. Pop, and P. Knabner, *Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation*, SIAM J. Numer. Anal., 42 (2004), pp. 1452–1478, https://doi.org/10.1137/S0036142902405229.

[121] F. A. Radu and W. Wang, *Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media*, Nonlin. Anal.: Real World Appl., 15 (2014), pp. 266–275, https://doi.org/10.1016/j.nonrwa.2011.05.003.

[122] L. A. Richards, *Capillary conduction of liquids through porous mediums*, Physics, 1 (1931), pp. 318–333, https://doi.org/10.1063/1.1745010.

[123] B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*, Frontiers in Applied Mathematics, SIAM, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717440.

[124] E. Saadatpoor, S. Bryant, and K. Sepehrnoori, *New trapping mechanism in carbon sequestration*, Transp. Porous Media, 82 (2010), pp. 3–17, https://doi.org/10.1007/s11242-009-9446-6.

[125] Schlumberger, *ECLIPSE: Technical Description 2013.2*, 2013, pp. 391–391, https://www.bibsonomy.org/bibtex/204b31cc0c9508c575df633b2a76d5c66/einar90.

[126] M. Schneider, L. Agélas, G. Enchéry, and B. Flemisch, *Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes*, J. Comput. Phys., 351 (2017), pp. 80–107, https://doi.org/10.1016/j.jcp.2017.09.003.

[127] D. Seus, K. Mitra, I. S. Pop, F. A. Radu, and C. Rohde, *A linear domain decomposition method for partially saturated flow in porous media*, Comput. Meth. Appl. Mech. Engrg, 333 (2018), pp. 331–355, https://doi.org/10.1016/j.cma.2018.01.029.

[128] M. Slodicka, *A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media*, SIAM J. Sci. Comp., 23 (2002), pp. 1593–1614, https://doi.org/10.1137/S1064827500381860.

[129] E. Spiteri, R. Juanes, M. J. Blunt, and F. M. Orr, *Relative-permeability hysteresis: Trapping models and application to geological CO_2 sequestration*, in SPE Annual Technical Conference and Exhibition, Dallas, Texas, 10 2005, https://doi.org/10.2118/96448-MS.

[130] D. Svyatskiy and K. Lipnikov, *Second-order accurate finite volume schemes with the discrete maximum principle for solving Richards' equation on unstructured meshes*, Adv. Water Resour., 104 (2017), pp. 114–126, https://doi.org/10.1016/j.advwatres.2017.03.015.

[131] T. A. Torp and J. Gale, *Demonstrating storage of CO_2 in geological reservoirs: The Sleipner and SACS projects*, Energy, 29 (2004), pp. 1361–1369, https://doi.org/10.1016/j.energy.2004.03.104.

[132] C. J. van Duijn, J. Molenaar, and M. J. de Neef, *The effect of capillary forces on immiscible two-phase flows in heterogeneous porous media*, Transp. Porous Media, 21 (1995), pp. 71–93, https://doi.org/10.1007/BF00615335.

[133] M. T. van Genuchten, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Sci. Soc. Amer. J., 44 (1980), pp. 892–898, https://doi.org/10.2136/sssaj1980.03615995004400050002x.

[134] V. Vilarrasa, D. Bolster, S. Olivella, and J. Carrera, *Coupled hydromechanical modeling of CO_2 sequestration in deep saline aquifers*, Int. Greenh. Gas Control, 4 (2010), pp. 910–919, https://doi.org/10.1016/j.ijggc.2010.06.006.

[135] X. Wang and H. A. Tchelepi, *Trust-region based solver for nonlinear transport in heterogeneous porous media*, J. Comput. Phys., 253 (2013), pp. 114–137, https://doi.org/10.1016/j.jcp.2013.06.041.

[136] A. W. Woods and A. Farcas, *Capillary entry pressure and the leakage of gravity currents through a sloping layered permeable rock*, J. Fluid Mech., 618 (2009), pp. 361–379, https://doi.org/10.1017/S0022112008004527.

[137] C. S. Woodward and C. N. Dawson, *Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media*, SIAM J. Numer. Anal., 37 (2000), pp. 701–724, https://doi.org/10.1137/S0036142996311040.

[138] Q. Zhou, J. T. Birkholzer, C.-F. Tsang, and J. Rutqvist, *A method for quick assessment of CO_2 storage capacity in closed and semi-closed saline formations*, Int. J. Greenh. Gas Control, 2 (2008), pp. 626–639, https://doi.org/10.1016/j.ijggc.2008.02.004.